

08/834734
PCT/EP

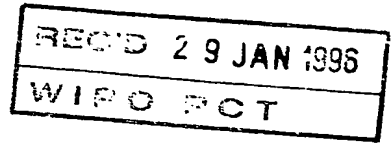
95/04575



Europäisches
Patentamt

European
Patent Office

Office européen
des brevets



Bescheinigung

Certificate

Attestation

Die angehefteten Unterla-
gen stimmen mit der
ursprünglich eingereichten
Fassung der auf dem näch-
sten Blatt bezeichneten
europäischen Patentanmel-
dung überein.

The attached documents
are exact copies of the
European patent application
described on the following
page, as originally filed.

Les documents fixés à
cette attestation sont
conformes à la version
initialement déposée de
la demande de brevet
européen spécifiée à la
page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

94402668.1

PRIORITY DOCUMENT

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

H. B. m.

H. Raaphorst

Den Haag, den
The Hague,
La Haye, le

15/01/96



Europäisches
Patentamt

European
Patent Office

Office européen
des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.:
Application no.:
Demande n°: **94402668.1**

Anmeldetag:
Date of filing:
Date de dépôt: **22/11/94**

Anmelder:
Applicant(s):
Demandeur(s):
ASSOCIATION FRANCAISE CONTRE LES MYOPATHIES
F-75013 Paris
FRANCE

Bezeichnung der Erfindung:
Title of the invention:
Titre de l'invention:
LGMD gene

In Anspruch genommene Priorität(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat:
State:
Pays:

Tag:
Date:
Date:

Aktenzeichen:
File no.
Numéro de dépôt:

Internationale Patentklassifikation:
International Patent classification:
Classification internationale des brevets:

C12N15/57, C12N9/64, A61K38/48, G01N33/50, C12Q1/68, C12N5/10, A61K48/00

Am Anmeldetag benannte Vertragsstaaten:
Contracting states designated at date of filing: **AT/BE/CH/DE/DK/ES/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE**
Etats contractants désignés lors du dépôt:

Bemerkungen:
Remarks:
Remarques:

1
LGMD gene

The invention relates to the isolated gene coding for a calcium dependent protease belonging to the Calpain family which, when it is mutated, is a cause of a disease called Limb-Girdle Muscular Dystrophy (LGMD).

The term limb-girdle muscular dystrophy (LGMD) was first proposed by Walton and Nattrass (1954) as part of a classification of muscular dystrophies. LGMD is characterised by progressive symmetrical atrophy and weakness of the proximal limb muscles and by elevated serum creatine kinase. Muscle biopsies demonstrate dystrophic lesions and electromyograms show myopathic features. The symptoms usually begin during the first two decades of life and the disease gradually worsens, often resulting in loss of walking ability 10 or 20 years after onset (Bushby, 1994). Yet, the precise nosological definition of LGMD still remains unclear. Consequently, various neuromuscular diseases such as facioscapulohumeral, Becker muscular dystrophies and especially spinal muscular atrophies have been occasionally classified under this diagnosis. For example, a recent study (Arikawa et al., 1991) reported that 17% (out of 41) of LGMD patients showed a dystrophinopathy. These issues highlight the difficulty in undertaking an analysis of the molecular and genetic defect(s) involved in this pathology.

Attempts to identify the genetic basis of this disease go back over 35 years. Morton and Chung (1959) estimated that "the frequency of heterozygous carrier ... is 16 per thousand persons". The same authors also stated that "the segregation analysis gives no evidence on whether these genes in different families are allelic or at different loci". Both autosomal dominant and recessive transmission have been reported, the latter being more common with an estimated prevalence of 10^{-5} (Emery, 1991). The localisation of a gene for a recessive form on chromosome 15 (LGMD2A, MIM 253600; Beckmann et al., 1991) provided the definitive proof that LGMD is a specific genetic entity. Subsequent genetic analyses confirmed this chromosome 15 localisation (Young et al., 1992; Passos-Bueno et al., 1993), the latter group demonstrating genetic heterogeneity of this disease. Although a recent study localised a second mutant

gene to chromosome 2 (LGMD2B, MIM 253601; Bashir et al., 1994), there is evidence that at least one other locus can be involved.

Genetic analyses of the LGMD2 kindreds revealed unexpected findings. First genetic heterogeneity was demonstrated in the highly inbred Indiana Amish community. Second although the Isle of la Réunion families were thought to represent a genetic isolate, at least 6 different disease haplotypes were observed, providing evidence against the hypothesis of a single founder effect (Beckmann et al., 1991) in this inbred population.

The nonspecific nosological definition, the relatively low prevalence and genetic heterogeneity of this disorder limit the number of families which can be used to restrict the genetic boundaries of the LGMD2A interval. Cytogenetic abnormalities, which could have helped to focus on a particular region, have not been reported. Immunogenetic studies of dystrophin-associated proteins (Matsumura et al., 1993) and cytoskeletal or extracellular matrix proteins such as a merosin (Tomé et al., 1994) failed to demonstrate any deficiency. In addition, there is no known specific physiological feature or animal model that could help to identify a candidate gene. Thus, there is no alternative to a positional cloning strategy.

It is established that the LGMD2 chromosomal region is localized on chromosome 15 as 15q15.1 - 15q21.1 region (Fougerousse et al., 1994).

Construction and analysis of a 10-12 Mb YAC contig (Fougerousse et al., 1994) permitted the mapping of 33 polymorphic markers within this interval and to further narrow the LGMD2A region to between D15S514 and D15S222. Furthermore, extensive analysis of linkage disequilibrium suggested a likely position for the gene in the proximal part of the contig.

The invention results from the construction of a partial cosmid map and the screening by cDNA selection (Lovett et al., 1991 ; Tagle et al., 1993) for muscle-expressed sequences encoded by this interval led to the identification of a number of potential candidate genes. One of these, previously cloned by Sorimachi et al. (1989), encodes a muscle specific protein, nCL1 (novel Calpain Large subunit 1), which belongs to the calpain family (CANP, calcium-activated neutral protease; EC 3.4.22.17), and appeared to be a functional candidate gene for this disease.

Calpains are non-lysosomal intracellular cysteine proteases which require calcium for their catalytic activities (for a review see Croall D.E. et al, 1991). The mammalian calpains include two ubiquitous proteins CANP1 and CANP2 as well as tissue-specific proteins. In addition to the muscle specific nCL1, stomach specific nCL2 and nCL2' proteins have also been described; these are derived from the same gene by alternative splicing. The ubiquitous enzymes consist of heterodimers with distinct large subunits associated with an common small subunit ; the association of tissue-specific large subunits with a small subunit has not yet been demonstrated. The large subunits of calpains can be subdivided into 4 protein domains. Domains I and III, whose functions remain unknown, show no homology with known proteins. Domain I, however, seems important for the regulation of the proteolytic activity. Domain II shows similarity with other cysteine proteases, sharing histidine, cysteine and asparagine residues at its active sites. Domain IV comprises four EF-hand structures which are potential calcium binding sites. In addition, three unique regions with no known homology are present in the muscle-specific nCL1 protein, namely NS, IS1 and IS2, the latter containing a nuclear translocation signal. These regions may be important for the muscle specific function of nCL1.

It is usually accepted that muscular dystrophies are associated with excess or deregulated calpains, and all the known approaches for curing these diseases are the use of antagonists of these proteases ; examples are disclosed in EP 359309 or EP 525420.

The invention results from the finding that, on the opposite to all these hypothesis, the LGMD2 disease is strongly correlated to the defect of a calpain which is expressed in healthy people.

The invention relates to the nucleic acid sequence such as represented in Figure 2 coding for a Ca^{++} dependent protease, or calpain, which is involved in LGMD2 disease, and more precisely LGMD2A. It also relates to a part of this sequence provided it is able to code for a protein having a calcium-dependent protease activity involved in LGMD2, or a sequence derived from one of the above sequences by substitution, deletion or addition of one or more nucleotides provided that said sequence is still coding for said protein, all the nucleic acids yielding a sequence complementary to a sequence as defined above.

The genomic organisation of the human nCL1 gene has been determined by the inventors, and consists of 24 exons and extends over 40 kb as represented in Figure 8, and is also a part of the invention. About 35 kb of this gene have been sequenced. A systematic screening of this gene in LGMD2A families led to the identification of 14 different mutations, establishing that a number of independent mutational events in nCL1 are responsible for LGMD2A. Furthermore, this is the first demonstration of a muscular dystrophy resulting from an enzymatic rather than a structural defect.

In the present specification, CANP3 means the protein which is a Ca^{++} dependent protease, or calpain, and coded by the nCL1 gene on chromosome 15.

The invention relates also to a protein, called CANP3, consisting in the amino acid sequence such as represented in figure 2 and which is involved, when mutated, in the LGMD2 disease.

The cDNA of the gene coding for CANP3, which is coding for the protein, is also represented in Figure 2, and is a part of the invention.

The protein coded by this DNA is CANP3, a calcium-dependent protease belonging to the Calpain family.

Are also included in the present invention the nucleic acid sequences derived from the cDNA of Figure 2 by one or more substitutions, deletions, insertions, or by mutations in 5' or 3' non coding regions or in splice sites, provided that the translated protein has the protease, calcium-dependent activity, and when mutated, induce LGMD2 disease.

The nucleic acid sequence encoding the protein might be DNA or RNA and be complementary to the nucleic acid sequence represented in Figure 2.

The invention also relates to a recombinant vector including a DNA sequence of the invention, under the control of a promoter allowing the expression of the calpain in an appropriate host cell.

A procaryotic or eucaryotic host cell transformed by or transfected with a DNA sequence comprising all or part of the sequence of Figure 2 is a part of the invention.

Such a host cell might be either :

- a cell which is able to secrete the protein and, this recombinant protein might be used as a drug to treat the LGMD2, or
 - a packaging cell line transfected by a viral or retroviral vector ; the cell lines bearing recombinant vector might be used as a drug for gene therapy of
- 5 LGMD2.

All the systems used today for gene therapy including adenoviruses and retroviruses and others described for example in « l'ADN médicament », (John Libbey, Eurotext, 1993), and bearing one of the DNA sequence of the invention are included herein by reference.

- 10 The examples hereunder and attached figures indicate how the structure of the gene was established, and how relationship between the gene and the LGMD was established.

Legend of the figures :

15 Figure 1:

A) Genomic organisation of the nCL1 gene

- The gene covers a 40 kb region of which 35 were sequenced (Accession number pending). Introns and exons are drawn to scale, the latter being indicated by numbered vertical bars. The first intron is the largest one and
- 20 remains to be fully sequenced. Position of intragenic microsatellites are indicated by asterisks. Arrows indicate the orientation of Alu (closed) and of Mer2 (greyed) repeat sequences.

B) *EcoRI* restriction map

- An *EcoRI* (E) restriction map of this region was established with the help
- 25 of cosmids from this region. The location of nCL1 gene is indicated as a black bar. The size of the corresponding fragments are indicated and are underlined when determined by sequence analysis.

C) Cosmid map of the nCL1 gene region.

- Cosmids were from a cosmid library constructed by subcloning YAC
- 30 774G4 (Richard in preparation) and are presented as lines. Dots on lines indicate positive STSs (indicated in boxed rectangles). A minimum of three cosmids cover the entire gene. T3,T7

Figure 2: Sequence of the human nCL1 cDNA (B) , and the flanking 5' (A) and 3' (C) genomic regions.

A) and C) The polyadenylation signal and putative CAAT, TATAA sites are boxed. Putative Sp1 (position -477 to -472), MEF2 binding sites (-364 to -343) and CArG box (-685 to -672) are in bold. The Alu sequence present in the 5' region is underlined.

B) The corresponding amino acids are shown below the sequence. The coding sequence between the ATG initiation codon and the TGA stop codon is 2466 bp, encoding for a 821 amino acid protein. The adenine in the first methionine codon has been assigned position 1. Locations of introns within the nCL1 gene are indicated by arrowheads. Nucleotides which differ from the previously published ones are indicated by asterisks.

Figure 3: Alignments of amino acid sequences of the muscle-specific calpains.

The human nCL1 protein is shown on the first line. The 3 muscle-specific sequences (NS, IS1 and IS2) are underlined. The second line corresponds to the rat sequence (Accession no P). The third and fourth lines show the deduced amino acid sequences encoded by pig and bovine Expressed Sequences Tagged (GenBank accession no U05678 and no U07858, respectively). The amino acids residues which are conserved among all known members of the calpains are in reverse letters. A period indicates that the same amino acid is present in the sequence. Letters refer to the variant amino acid found in the homologous sequence. Position of missense mutations are given as numbers above the mutated amino acid.

Figure 4: Distribution of the mutations along nCL1 protein structure.

A) Positions of the 23 introns are indicated by vertical bars in relation to the corresponding amino acid coordinates.

B) The nCL1 protein is depicted showing the four domains (I, II, III, IV) and the muscle specific sequences (NS, IS1 and IS2). The position of missense mutations within nCL1 domain are indicated by black dots. The effect of nonsense and frameshift mutations are illustrated as truncated lines, representing the extent of protein synthesised. Name of the corresponding families are indicated on the left of the line. The out of frame ORF is given by hatched lines.

Figure 5: Northern blot hybridisation of a nCL1 clone

A mRNA blot (Clontech) containing 2 µg of poly(A)+ RNA from each of eight human tissues was hybridised with a nCL1 genomic clone spanning exons 20 and 21. The latter detects a 3.6 kb mRNA present only in a line
5 corresponding to the skeletal muscle mRNA.

Figure 6: Representative mutations identified by heteroduplex analysis.

Examples of mutation screening by heteroduplex analysis. Pedigree B505 shows the segregation of two different mutations in exon 22.

Figure 7: Homozygous mutations in the nCL1 gene

10 Detection by sequencing of mutations in exons 2 (a), 8 (b), 13 (c) and 22 (d). Sequences from a healthy control are shown above each mutant sequence. Asterisks indicate the position of the mutated nucleotides. The consequences on codon and amino acid residues are indicated on the left of the figure together with the name of the family.

15 **Figure 8:** Structure of nCL1 gene

Figure 8A represents the 5' part of the gene with exon 1.

Figure 8B represents the part of the gene including exons 2 to 8,

Figure 8C represents the part of the gene including exon 9,

Figure 8D represents the part of the gene including exons 10 to 24
20 including the 3' non transcribed region.

EXAMPLES

EXAMPLE 1

Localisation of the nCL1 within the LGMD2A interval

Detailed genetic and physical maps of the LGMD2A region were
25 constructed (Fougerousse et al., 1994), following the primary linkage assignment to 15q (Beckmann et al., 1991). The disease locus was bracketed between the D15S129 and D15S143 markers, defining the cytogenetic boundaries of the LGMD2A region as 15q15.1-15q21.1 (Fougerousse et al., 1994). Construction and analysis of a 10-12 Mb YAC contig (Fougerousse et al., 1994) permitted us
30 to map 33 polymorphic markers within this interval and to further narrow the LGMD2A region to between D15S514 and D15S222.

The nCL1 gene had been localised to chromosome 15 by hybridisation with sorted chromosomes and by Southern hybridisation to DNA from human-mouse cell hybrids (Ohno et al., 1989). cDNA capture using YACs from the LGMD2A interval allowed the identification of thirteen positional candidate genes. nCL1 was one of the two transcripts identified that showed muscle-specific expression as evidenced by northern blot analysis. The localisation was further confirmed by STS (for Sequence Tagged Site) assays. Primers used for the localisation of the nCL1 gene are P94in2, P94in13 and pcr6a3, as shown in Figure 1 and their characteristics being defined in Table 1.

Table 1: PCR primers used for localisation of the nCL1 gene.

Primer name	Primer sequence (5'-3')	Position within the cDNA	Annealing temp (°C)	PCR product size on	
				cDNA	genomic DNA
P94in2	ATGGAGCCAACAGAACTGA C GTATGACTCGGAAAAGAAG GT	341-360 428-448	58	108	1758
P94in13	TAAGCAAAAGCAGTCCCCA C TTGCTGTTCTCACTTTCTCCT G	1893-1912 1936-1956	58	64	1043
P94-6a3	GTTTCATCTGCTGCTTCGTT CTGGTTCAGGCATACATGG T	2342-2361 2452-2471	56	130	818
P94ex1ter	TTCTTTATGTGGACCCTGAG TT ACGAACTGGATGGGGAACT	218-239 275-293	55	76	76

These primers are designed from different parts of the published human cDNA sequence (Sorimachi et al., 1989), and were used for an STS content screening on DNA from three chromosome 15 somatic cell hybrids and YACs from the LGMD2A contig. The results positioned the gene in a region previously defined as 15q15.1-q21.1 and on 3 YACs (774G4, 926G10, 923G7) localised in this region. The relative positions of STSs along the LGMD2A contig allowed to localise the gene between D15S512 and D15S488, in a candidate region suggested by linkage disequilibrium studies.

The same primers as above were used to screen a cosmid library from YAC 774G4. A group of 5 cosmids was identified (Fig. 1). Experiments with another nCL1 primer pair (P94ex1ter; Table 1) established that these cosmids cover all nCL1 exons except number 1, and that a second group of 4 cosmids contain this

exon (Fig. 1). A minimal set of three overlapping cosmids (2G8-2B11-1F11) covers the entire gene (Figure 1). DNA from these cosmids was used to construct an *EcoRI* restriction map of this region (Figure 1B).

EXAMPLE 2

5 **Determination of the nCL1 gene sequence**

Most of the sequences were obtained through shotgun sequencing of partial digests of cosmid 1F11 subcloned in M13 and bluescript vectors, and by walking with internal primers. The sequence assembly was made using the XBAP software of the Staden package (Staden) and was in agreement with the
10 restriction map of the cosmids. Sequences of exon 1 and adjacent regions were obtained by sequencing cosmid DNA or PCR products from human genomic DNA. The first intron is still not fully sequenced, but there is evidence that it may be between 10 to 16 kb in length (based on hybridisation of restriction fragments; data not shown). The entire gene, including its 5' and 3' regions, is more than 40
15 kb long, and shown in Figure 8.

a) the cDNA sequence

The used technology allows the implementation of the published human cDNA sequence of nCL1 (Sorimachi 1989). It contains the missing 129 bases corresponding to the N-terminal 43 amino acids (Figure 2). It also differs from it
20 at 12 positions. Three of which occur at third base positions of codons and preserve the encoded amino acid sequence. The other 9 differences lead to changes in amino-acid composition (Figure 2). As these different exons were sequenced repeatedly on at least 10 distinct genomes, we are confident that the sequence of Fig. 2 represents an authentic sequence and does not contain
25 minor polymorphic variants. Furthermore, these modifications increase the local similarity with the rat nCL1 amino acid sequence (Sorimachi), although the overall similarity is still 94 %.

The ATG numbered 1 in Figure 2 is the translation initiation site based on homology with the rat nCL1, and is within a sequence with only 5 nucleotides out
30 of 8 in common with the Kosak consensus sequence (Kosak M, 1984). Putative CCAAT and TATA boxes were observed 590, 324, (CCAAT) and 544 or 33 bp (TATA) upstream of the initiating ATG codon, respectively (Bucher, 1990). A GC-box binding the Sp1 protein (Dyran et al., 1983) was identified at position -477.

Consensus sequences corresponding to potential muscle-specific regulatory elements were identified (Fig. 2). These include a myocyte-specific enhancer-binding factor 2 (MEF2) binding site (Cserjesi P. 1991), a CArG box (Minty A. 1986) and 6 E-boxes (binding sites for basic Helix-Loop-Helix proteins frequently found in members of MyoD family; Blackwell et Weintraub, 1990). The functional significance of these putative transcription factor binding sites in the regulation of nCL1 gene expression remains to be established.

Two potential AAUAAA polyadenylation signals, were identified 520 and 777 bp downstream of the TGA stop codon. The sequencing of a partial nCL1 cDNA containing a polyA tail, demonstrated that the first AAUAAA is the polyadenylation signal. The latter is embedded in a region well conserved with the rat nCL1 sequence and is followed after 4 bp by a G/T cluster, present in most genes 3' of the polyadenylation site (Birnstiel et al., 1985). The 3'-untranslated region of the nCL1 mRNA is 565 bp long. The predicted length of the cDNA should therefore be approximately 3550 or 3000 bp.

b) Comparison with calpain

The sequence of the human nCL1 gene was compared to those of other calpains thereof (Figure 3). The most telling comparisons are with the homologous rat (Accession no J05121), bovine (Accession no U07858) and porcine (Accession no U05678) sequences. The accession numbers refers to those or international genebanks, such as GeneBank (N.I.H.) or EMBL Database (EMBL, Heidelberg). High local similarities between the human and rat DNA sequences are even observed in the 5' (75%) or in different parts of the 3' untranslated regions (over 60%) (data not shown). The high extent of sequence homology manifested by the human and rat nCL1 gene in their untranslated regions is suggestive of evolutionary pressures on common putative regulatory sequences.

c) Genomic organisation of the nCL1 gene

A comparison of the published nCL1 human cDNA (Sorimachi et al., 1989) with the corresponding genomic sequence led to the identification of 24 exons ranging in length from 12 bp (exon 13) to 309 bp (exon 1), with a mean size of 100 bp (Figure 1). The size of introns ranges from 86 bp to about 10-16 kb for intron 1.

The intron-exon boundaries as shown in Table 2 exhibit close adherence to 5' and 3' splice site consensus sequences (Shapiro and Senapathy, 1987).

Table 2: Sequences at the intron-exon junctions. A score expressing adherence to the consensus was calculated for each site according to Shapiro and Senapathy (1987). Sequences of exons and introns are in upper and lower cases, respectively. Size of exons are given in parenthesis.

splice donor site	score (%)	Intron	score (%)	splice acceptor site	Exon
...CTCCGgtgagt...	88.5	<-Intron 1->	99.0	...ttttgtttcacagGAAAT...	Exon 1 (309 bp) ->
...GCTAGgtagga...	83.5	<-Intron 2->	90.0	...gtgtctgcctgcagGGGAC...	Exon 2 (70 bp) ->
...TCCAGgtgagg...	92	<-Intron 3->	81.5	...acgcttctgtgcagTTCTG...	Exon 3 (119 bp) ->
...GCTAAgtaagc...	82	<-Intron 4->	81.5	...atcctctctctaagGCTCC...	Exon 4 (134 bp) ->
...TTGATgtaagt...	87	<-Intron 5->	79.5	...ccatcgggcctcagGATGG...	Exon 5 (169 bp) ->
...CCCGGgtgtgt...	77.5	<-Intron 6->	91	...ttactgtctctacagACAAT...	Exon 6 (144 bp) ->
...ATGAGgtaagc...	94	<-Intron 7->	78.5	...tctgtgtgccttaagGTCCC...	Exon 7 (84 bp) ->
...GATAGgtaggt...	89	<-Intron 8->	91.5	...cattttcccaccagATGGA...	Exon 8 (86 bp) ->
...TTCTGgtgagt...	88	<-Intron 9->	92	...ttccaacctctcagGATGT...	Exon 9 (78 bp) ->
...CCCAGgtggga...	80	<-Intron 10->	68.5	...ttctgggggtgcagATACT...	Exon 10 (161 bp) ->
...ACGAGgtgtgt...	85.5	<-Intron 11->	86	...tgtttcttctcaagGTTCC...	Exon 11 (170 bp) ->
...AAGAGgtatag...	70	<-Intron 12->	87	...tccccatctctcagATGCA...	Exon 12 (12 bp) ->
...TCTGAgtgagt...	76.5	<-Intron 13->	97	...tgtattcctcacagGGAAG...	Exon 13 (209 bp) ->
...CAGTGgtgagt...	89	<-Intron 14->	93.5	...cttttctatgcagAAAAA...	Exon 14 (37 bp) ->
...CCAAGgtaggt...	89	<-Intron 15->	87	...cctcctctctccagCCCAT...	Exon 15 (18 bp) ->
...CACAGgtgtct...	80	<-Intron 16->	88	...ttgtgcctccacagCCACA...	Exon16 (114 bp) ->
...GAGATgtgagt...	84	<-Intron 17->	92.5	...cccttctcctcagGACAT...	Exon 17 (78 bp) ->
...CAAACgtgagt...	83	<-Intron 18->	90	...ctccatccccccagACAAG...	Exon 18 (58 bp) ->
...TGGATgtatcc...	56	<-Intron 19->	88	...cctccctctccagACAGA...	Exon 19 (65 bp) ->
...GGCAGgtggga...	80	<-Intron 20->	94	...ttttctattgccagAAATA...	Exon 20 (69 bp) ->
...CGCAGgtgctg...	66	<-Intron 21->	91	...gggtcccctccacagGATTC...	Exon 21 (79 bp) ->
					Exon 22 (117 bp) ->

...GTTCAGtaagt...	79	<-Intron 22->	93.5	...gcattctttcacagGAGCT...	Exon 23 (59 bp) ->
...TGGAGgtaaag...	81	<-Intron 23->	79	...gggacttctttcagTGGCT...	Exon 24 (27 bp) ->

When the genomic sequence was submitted to GRAIL analysis (Uberbacher et al., 1991), 11 exons were correctly recognised, 4 were not identified, 6 were inadequately defined and 2 were too small to be recognised (data not shown).

5 As already noted, the nCL1 gene has three unique sequence blocks, NS (amino acid residues 1 to 61), IS1 (residues 267 to 329) and IS2 (residues 578 to 653). It is interesting to note that each of these sequences, as well as the nuclear translocation signal inside IS2, are essentially flanked by introns (Fig. 4). The exon-intron organisation of the human nCL1 is similar to that reported for
10 the chicken CANP (the only other large subunit calpain gene whose genomic structure is known; (Emori et al., 1986).

Four microsatellite sequences were identified. Two of them are in the distal part of the first intron: an (AT)₁₄ and an previously identified mixed-pattern microsatellite, S774G4B8, which was demonstrated to be non polymorphic
15 (Fougerousse et al., 1994). A (TA)₇(CA)₄(GA)₁₃ was identified in the second intron and genotyping of 64 CEPH unrelated individuals revealed two alleles (with frequencies of 0.10 and 0.90). The fourth microsatellite is a mixed (CA)_n(TA)_m repeat present in the 9th intron. The latter and the (AT)₁₄ repeat have not been investigated for polymorphism. Fourteen repetitive sequences of
20 the Alu family and one Mer2 repeat were identified in the nCL1 gene (Fig. 1C), which has, thus, on the average one Alu element per 2.5 kb.

Southern blot experiments (Ohno et al., 1989) and STS screening (data not shown) suggest that there is but one copy per genome of this member of the calpain family.

25 EXAMPLE 3

Expression of the nCL1 gene

The pattern of tissue-specificity was investigated by northern blot hybridisation with a genomic subclone probe from cosmid 1F11 spanning exons 20 and 21. There is no evidence for the existence of an alternatively spliced form
30 of nCL1, although this cannot be excluded. A transcript of about 3.4-3.6 kb was

detected in skeletal muscle mRNA (Figure 5). This size therefore favours that the position -544 is the functional TATA box.

Transcription studies suggested that it is an active gene rather than a pseudogene and its muscle-specific pattern of expression is consistent with the phenotype of this disorder (Sorimachi et al., 1989 and Figure 5).

EXAMPLE 4

Mutation screening

nCL1 fulfils both positional and functional criteria to be a candidate gene for LGMD2A. To evaluate its role in the etiology of this disorder, nCL1 was systematically screened in 38 LGMD2 families for the presence of nucleotide changes using a combination of heteroduplex (Keen et al., 1991) and direct sequence analyses.

PCR primers were designed to specifically amplify the exons and splice junctions and also the regions containing the putative CAT, TATA boxes and the polyadenylation signal of the gene as shown in Table 3.

Table 3: PCR primers used for the analysis of the nCL1 gene in LGMD patients.

amplified region	Primer sequences (5'-3')	Size (bp)	Annealing temp. (°C)
promotor	TTCAGTACCTCCCGTTCACC GATGCTTGAGCCAGGAAAAC	296	59
exon 1	CTTTCCTTGAAGGTAGCTGTAT GAGGTGCTGAGTGAGAGGAC	438	60
exon 2	ACTCCGTCTCAAAAAAATACCT ATTGTCCCTTTACCTCCTGG	239	57
exon 3	TGGAAGTAGGAGAGTGGGCA GGGTAGATGGGTGGGAAGTT	354	58
exon 4	GAGGAATGTGGAGGAAGGAC TTCCTGTGAGTGAGGTCTCG	292	59
exon 5	GGAACCTCTGTGACCCCAAAT TCCTCAAACAAAACATTCGC	325	56
exon 6	GTTCCCTACATTCTCCATCG GTTATTTCAACCCAGACCCTT	315	57
exon 7	AATGGGTTCTCTGGTTACTGC AGCACGAAAAGCAAAGATAAA	333	56
exon 8	GTAAGAGATTGCCCCCAG TCTGCGGATCATTGGTTTGT	321	58
exon 9	CCTTCCCTTCTTCTGCTTC CTCTCTTCCCCACCCTTACC	173	56
exon 10	CCTCCTCACCTGCTCCCAT TTTTTCGGCTTAGACCCTCC	251	56
exon 11	TGTGGGGAATAGAAATAAATGG CCAGGAGCTCTGTGGGTCA	355	57
exon 12	GGCTCCTCATCCTCATTCACA GTGGAGGAGGGTGAGTGTGC	312	61
exon 13	TGTGGCAGGACAGGACGTTC	337	60

	TTCAACCTCTGGAGTGGGCC		
exon 14	CACCAGAGCAAACCGTCCAC	230	61
	ACAGCCCAGACTCCCATTC		
exon 15	TTCTCTTCTCCCTTCACCCT	225	57
	ACACACTTCATGCTCTCTACCC		
exon 16	CCGCCTATTCTTTTCCTCTT	331	56
	GACAAACTCCTGGGAAGCCT		
exon 17	ACCTCTGACCCCTGTGAACC	270	61
	TGTGGATTTGTGTGCTACGC		
exon 18	CATAAATAGCACCGACAGGGA	258	59
	GGGATGGAGAAGAGTGAGGA		
exon 19	TCCTCACTCTTCTCCATCCC	159	57
	ACCCTGTATGTTGCCTTGG		
exons 20-21	GGGGATTTTGCTGTGTGCTG	333	61
	ATTCCTGCTCCCACCGTCTC		
exon 22	CACAGAGTGTCCGAGAGGCA	282	57
	GGAGATTATCAGGTGAGATGCC		
exons 22-23	CAGAGTGTCCGAGAGGCAGGG	608	61
	CGTTGACCCCTCCACCTTGA		
exon 24	GGGAAAACATGCACCTTCTT	375	58
	TAGGGGGTAAAATGGAGGAG		
polyadenylation signal	ACTAACTCAGTGGGAATAGGG	413	56
	GGAGCTAGGATAGCTCAAT		

PCR products made on DNA from blood of specific LGMD2A patients were then subjected either to heteroduplex analysis or to direct sequencing, depending on whether the mutation, based on haplotype analysis, was expected to be homozygous or heterozygous, respectively. It was occasionally necessary to clone the PCR products to precisely identify the mutations (i.e., for microdeletions or insertions and for some heterozygotes). Disease-associated mutations are summarised in Table 4 hereunder and their position along the protein is shown in Fig. 4.

Table 4: nCL1 mutations in LGMD2A families.

Codons and amino acid positions are numbered on the basis of the cDNA sequence starting from ATG.

Exon	Families	Nucleotide position	Nucleotide change	Amino acid position	Amino acid change	Restriction si
2	B519*	328	<u>C</u> GA-> <u>T</u> GA	110	Arg->stop	
4	M42	545	CT <u>G</u> -> C <u>A</u> G	182	Leu->Gln	
4	M1394; M2888	550	CAA -> CA	184	frameshift	
5	M35; M37	701	GG <u>G</u> -> G <u>A</u> G	234	Gly->Glu	

6	M32	945	CGG -> CG	315	frameshift	-SmaI
8	M2407*	1061	G <u>T</u> G -> G <u>G</u> G	354	Val-> Gly	
8	M1394	1079	T <u>G</u> G -> T <u>A</u> G	360	Trp->stop	-BstnI, -Eco
11	M2888	1468	<u>C</u> GG -> <u>T</u> GG	490	Arg->Trp	
13	R12*	1715	<u>C</u> GG -> <u>C</u> AG	572	Arg->Gln	-MspI
19	R27	2069-2070	deletion AC	690	frameshift	
21	R14; R17	2230	<u>A</u> GC -> <u>G</u> GC	744	Ser->Gly	-AluI
22	A*; B501*; M32	2306	<u>C</u> GG -> <u>C</u> AG	769	Arg->Gln	
22	B505	2313-2316	deletion AGAC	771-772	frameshift	
22	R14; B505	2362-2363	AG -> TCATCT	788	frameshift	

The first letter of the family code refers to the origin of the population B= Brazil, M= metropolitan France, R = Isle of La Réunion, A= Amish.

Each mutation was confirmed by heteroduplex analysis, by sequencing of both strands in several members of the family or by enzymatic digestion when the mutation resulted in the modification of a restriction site. Segregation analyses of the mutations, performed on DNAs from all available members of the families, confirmed that these sequence variations are on the parental chromosome carrying the LGMD2A mutation. To exclude the possibility that the missense substitutions might be polymorphisms, their presence was systematically tested in a control population: none of these mutations was seen among 120 control chromosomes from the CEPH reference families.

EXAMPLE 5 :

Analysis of families genes, chromosome-15 ascertained families

The initial screening for causative mutations was performed on families, each containing a LGMD gene located on chromosome 15. These included families from the Island of La Réunion (Beckmann et al., 1991), from the Old Order Amish from northern Indiana (Young et al., 1992,) and 2 Brazilian families (Passos Bueno et al., 1993).

a) Reunion Island families

Genealogical studies and geographic isolation of the families from the Isle of La Réunion were suggestive of a single founder effect. Genetic analyses are,

however, inconsistent with this hypothesis as the families present haplotype heterogeneity. At least, six different carrier chromosomes are encountered, (with affected individuals in several families being compound heterozygotes). Distinct mutations corresponding to four of these six haplotypes have been identified thus far.

In family R14, exons 13, 21 and 22 showed evidence for sequence variation upon heteroduplex analysis (Fig. 6). Sequencing of the associated PCR products revealed (i) a polymorphism in exon 13, (ii) a missense mutation (A->G) in exon 21 transforming the Ser⁷⁴⁴ residue to a glycine in the loop of the second EF-hand in domain IV of the protein (Figure 4), and (iii) a frameshift mutation in exon 22. The exon 21 mutation and the polymorphism in exon 13 form an haplotype which is also encountered in family R17. Subcloning of the PCR products was necessary to identify the exon 22 mutation. Sequencing of several clones revealed a replacement of AG by TCATCT (data not shown). This frameshift mutation causes premature termination at nucleotide 2400 where an in frame stop codon occurs (Figure. 4).

The affected individuals in family R12 are homozygous for all markers of the LGMD2A interval (Allamand, submitted). Sequencing of the PCR products of exon 13 revealed a G to A transition at base 1715 of the cDNA resulting in a substitution of glutamine for Arg⁵⁷² (Figure. 7) within domain III, a residue which is highly conserved throughout all known calpains. This mutation, detectable by loss of *MspI* restriction site, is present only in this family and in no other examined LGMD2A families or unrelated controls.

In family R27, heteroduplex analysis followed by sequencing of the PCR products of an affected child revealed a two base pair deletion in exon 19 (Figure. 6 and table 4). One AC out of three is missing at this position of the sequence, producing a stop codon at position 2069 of the cDNA sequence (Figure 4).

b) Amish families

As expected, due to multiple consanguineous links, the examined LGMD2A Northern Indiana Amish patients were homozygous for the haplotype on the chromosome bearing the mutant allele (Allamand, submitted). A (G->A) missense mutation was identified at nucleotide 2306 within exon 22 (Fig. 7). The

resulting codon change is CGG to CAG, transforming Arg⁷⁶⁹ to glutamine. This residue, which is conserved throughout all members of the calpain family in all species, is located in domain IV of the protein within the 3rd EF-hand at the helix-loop junction (ref). This mutation was encountered in a homozygous state
 5 in all patients from 12 chromosome 15-linked Amish families, in agreement with the haplotype analysis. We also screened six Southern Indiana Amish LGMD families, for which the chromosome 15 locus was excluded by linkage analyses (Allamand ESHG, submitted, ASHG 94). As expected, this nucleotide change was not present in any of the patients from these families, thus confirming the
 10 genetic heterogeneity of this disease in this genetically related isolate.

c) Brazilian families

As a result of consanguineous marriages, two Brazilian families (B501, B519) are homozygous for extended LGMD2A carrier haplotypes (data not shown). Sequencing PCR products from affected individuals of these families
 15 demonstrated that family B501 has the same exon 22 mutation found in northern Indiana Amish patients (Figure 7), but embedded in a completely different haplotype. In family B519, the patients carry a C to T transition in exon 2, replacing Arg³²⁸ with a TGA stop codon (Figure 7), thus leading, presumably, to a very truncated protein (Figure 4).

20 d) Analysis of other LGMD families

Having validated the role of the candidate gene in the chromosome 15 ascertained families, we next examined by heteroduplex analysis LGMD families for which linkage data were not informative. These included one Brazilian (B505) and 13 metropolitan French pedigrees.

25 Heteroduplex bands were revealed for exons 1, 3, 4, 5, 6, 8, 11, 22 of one or more patients (Figure 6). Of all sequence variants, 10 were identified as possible pathogenic mutations (5 missense, 1 nonsense and 4 frameshift mutations) and 3 as polymorphisms with no change of amino acid of the protein. All causative mutations identified are listed in Table 4 here-above. Identical
 30 mutations were uncovered in apparently unrelated families. The mutations shared by families M35 and M37, and M2888 and M1394, respectively, are likely to be the consequence of independent events since they are embedded in different marker haplotypes. In contrast, it is likely that the point mutation in exon

22 of the Amish and in the M32 kindreds corresponds to the same mutational event as both chromosomes share a common four marker haplotype (774G4A1-774G4A10-774G454D-774G4A2) around nCL1 (data not shown), possibly reflecting a common ancestor. The same holds true for the AG to TCATCT substitution mutation encountered in exon 22 in families B505 and R14. The exon 8 (T->G) transversion is present in the two carrier chromosomes of M2407, the only metropolitan family homozygous by haplotype, possibly reflecting an undocumented consanguinity. For some families, no disease-causing mutation has been detected thus far (M40 for example).

In addition to the polymorphism present in exon 13 in families R14 and R17 (position 668) and in the intragenic microsatellites, four additional neutral variations were detected: a (T->C) transition at position 96, abolishing a *DdeI* restriction site in exon 1 in M31; a (C->T) transition in exon 3 (position 495) in M40 and in M37 forming a haplotype with the exon 5 mutation (in the former family, this polymorphism does not cosegregate with the disease); a (T->C) transition in the paternally derived promotor in M42 at position -428, which was also evidenced in healthy controls; and a variable poly(G) in intron 22 close to the splice site in families R20, R11, R19, M35 and M37. The latter is also present in the members of the CEPH families, but is not useful as a genetic marker as the visualisation and interpretation of mononucleotide repeat alleles is difficult.

In total, sixteen independent mutational events representing fourteen different mutations were identified. All mutations cosegregate with the disease in LGMD2A families. The characterised morbid calpain alleles contain nucleotide changes which were not found in alleles from normal individual. The discovery of two nonsense and five frameshift mutations in nCL1 supports the hypothesis that a deficiency of this product causes LGMD2A. All seven mutations result in a premature in-frame stop codon, leading to the production of truncated and presumably inactive proteins (Figure 4). Evidences for the morbidity of the missense mutations come from (1) the relative high incidence of such mutations among LGMD2A patients ; although it is difficult in the absence of functional assays to differentiate between a polymorphism and a morbid mutation, the occurrence of different "missense" mutations in this gene cannot all be

accounted for as rare private polymorphisms; (2) the failure to observe these mutations in control chromosomes; and (3) the occurrence of mutations in evolutionarily conserved residues and/or in regions of documented functional importance. Four of seven missense mutations change an amino acid which is conserved in all known members of the calpain family in all species (Figure 3). Two of the remaining mutations affect less conserved amino acid residues, but are located in important functional domains. The substitution V354G in exon 8 is 4 residues before the asparagine at the active site and S744G in exon 21 is within the loop of the second EF-hand and may impair the calcium-dependent regulation of calpain activity or the interaction with a small subunit (Figure 4). Several missense mutations change a hydrophobic residue to a polar one, or vice versa (Table 4) possibly disrupting higher order structures.

METHODS

Description of the patients

The LGMD2A families analysed were from 4 different geographic origins. They included 3 Brazilian families, 13 interrelated nuclear families from the Isle of la Réunion, 10 French metropolitan families and 12 US Amish families. The majority of these families were previously ascertained to belong to the chromosome 15 group by linkage analysis (Beckmann, 1991; Young, Passos-Bueno et al., 1993). However, some families from metropolitan France as well as one Brazilian family, B505, had non significant lodscores for chromosome 15. Genomic DNA was obtained from peripheral blood lymphocytes.

Sequencing of cosmid c774G4-1F11 and EcoRI restriction map of cosmids.

Cosmid 1F11 (Figure 1C) was subcloned following DNA preparation through Qiagen procedure (Qiagen Inc., USA) and partial digestion with either *Sau3A*, *RsaI* or *AluI*. Size-selected restriction fragments were recovered from low-melting agarose and eventually ligated with M13 or Bluescript (Stratagene, USA) vectors. After electroporation in *E.coli*, recombinant colonies were picked in 100 µl of LB/ampicillin media. PCR reactions were performed on 1 µl of the culture in 10 mM Tris-HCl, pH 9.0, 50 mM KCl, 1.5 mM MgCl₂, 0.1% Triton X-100, 0.01 gelatine, 200µM of each dNTP, 1 U of Taq Polymerase (Amersham) with 100 ng of each vectors primers. Amplification was initiated by 5 min denaturation at 95°C, followed by 30 cycles of 40 sec denaturation at 92°C and 30 sec annealing

at 50°C. PCR products were purified through Microcon devices (Amicon, USA) and sequenced using the dideoxy chain termination method on an ABI sequencer (Applied Biosystems, Foster City, USA). The sequences were analysed and alignments performed using the XBAP software of the Staden package, version 93.9 (Staden, 1982). Gaps between sequence contigs were
5 filled by walking with internal primers. *EcoRI* restriction map of cosmids was performed essentially as described in Sambrook et al. (1989).

Northern Blot analysis

The probes were labelled by random priming with dCTP-($\alpha^{32}\text{P}$).
10 Hybridisation was performed to human multiple tissue northern blots as recommended by the manufacturer (Clontech, USA).

Analysis of PCR products from LGMD2A families

One hundred ng of human DNA were used per PCR under the buffer and cycle conditions described in Fougere (1994) (annealing temperature shown
15 in Table 3). Heteroduplex analysis (Keene et al., 1991) was performed by electrophoresis of ten μl of PCR products on a 1.5 mm-thick Hydrolink MDE gels (Bioprobe) at 500-600 volt for 12-15 h depending of the fragment length. Migration profile was visualised under UV after ethidium bromide staining.

For sequence analysis, the PCR products were subjected to dye-dideoxy
20 sequencing, after purification through microcon devices (Amicon, USA). When necessary, depending on the nature of the mutations (e.g., frameshift mutation or for some heterozygotes), the PCR products were cloned using the TA cloning kit from Invitrogen (UK). One μl of product was ligated to 25 ng of vector at 12°C overnight. After electroporation into XL1-blue bacteria, several independent
25 clones were analysed by PCR and sequenced as described above.

The invention results from the finding that the nCL1 gene when it is mutated is involved in the etiology of LGMD2A. It is exactly the contrary to what is stated in the literature, e.g. that the disease is accompanied by the presence of a deregulated calpain. Identification of nCL1 as the defective gene in LGMD2A
30 represents the first example of muscular dystrophy caused by mutation affecting a gene which is not a structural component of muscle tissue, in contrast with previously identified muscular dystrophies such as Duchenne and Becker (Bonilla et al., 1988), severe childhood autosomal recessive (Matsumara et al.,

1992), Fukuyama (Matsumara et al., 1993) and merosin-deficient congenital muscular dystrophies (Tomé et al., 1994).

The understanding of the LGMD2A phenotype needs to take into account the fact that there is no active nCL1 protein in several patients, a loss compatible with the recessive manifestation of this disease. Simple models in which this protease would be involved in the degradation or destabilisation of structural components of the cytoskeleton, extracellular matrix or dystrophin complex can therefore be ruled out. Furthermore, there are no signs of such alterations by immunocytogetic studies on LGMD2 muscle biopsies (Matsumara et al., 1993; Tomé et al., 1994). Likewise, since LGMD2A myofibers are apparently not different from other dystrophic ones, it seems unlikely that this calpain plays a role in myoblast fusion, as proposed for ubiquitous calpains (Wang et al., 1989).

All the data disclosed in these examples confirm that the nCL1 gene is a major gene involved in the disease when mutated.

The fact that morbidity results from the loss of an enzymatic activity raises hopes for novel pharmaco-therapeutic prospects. The availability of transgenic models will be an invaluable tool for these investigations.

The invention is also relative to the use of a nucleic acid or a sequence of nucleic acid of the invention, or to the use of a protein coded by the nucleic acid for the manufacturing of a drug in the prevention or treatment of LGMD2.

The finding that a defective calpain underlies the pathogenesis of LGMD2A may prove useful for the identification of the other loci involved in the LGMDs. Other forms of LGMD may indeed be caused by mutations in genes whose products are the CANP substrates or in genes involved in the regulation of nCL1 expression. Techniques such as the two-hybrid selection system (Fields et al., 1989) could lend themselves to the isolation of the natural protein substrate(s) of this calpain, and thus potentially help to identify other LGMD loci.

The invention also relates to the use of all or a part of the peptidic sequence of the enzyme, or of the enzyme, product of nCL1 gene, for the screening of the ligands of this enzyme, which might be also involved in the etiology and the morbidity of LGMD2

The ligands which might be involved are for example substrate(s), activators or inhibitors of the enzyme.

The nucleic acids of the invention might also be used in a screening method for the determination of the components which may act on the regulation of the gene expression.

5 A process of screening using either the enzyme or a host recombinant cell, containing the nCL1 gene and expressing the enzyme, is also a part of the invention.

The pharmacological methods, and the use of nucleic acid and peptidic sequences of the invention are very potent applications.

10 The methods used for such screenings of ligands or regulatory elements are those described for example for the screening of ligands using cloned receptors.

The identification of mutations in the nCL1 gene provides the means for direct prenatal or presymptomatic diagnosis and carrier detection in families in which both mutations have been identified. Gene-based accurate classification of LGMD2A families should prove useful for the differential diagnosis of this
15 disorder.

The invention relates to a method of detection of a predisposition to LGMD2 in a family or a human being, such method comprising the steps of :

- selecting one or more exons or flanking sequences which are sensitive in said family;
- 20 - selecting the primers specific for the or these exons or their flanking sequences, a specific example being the PCR primers of Table 3, or an hybrid thereof,
- amplifying the nucleic acid sequence, the substrate for this amplification being the DNA of the human being to be checked for the predisposition, and
- 25 - comparing the amplified sequence to the corresponding sequence derived from Figure 2 or Figure 8.

Table 2 indicates the sequences of the introns-exons junctions, and primers comprising in their structure these junctions are also included in the invention.

30 All other primers suitable for such RNA or DNA amplification may be used in the method of the invention.

In the same way, any suitable amplification method : PCR (for Polymerase Chain Reaction ®) NASBA ® (for Nucleic acid Sequence Based Amplification), or others might be used.

The methods usually used in the detection of one site mutations, like ASO (Allele specific PCR), LCR, or ARMS (Amplification Refactory Mutation System) may be implemented with the specific primers of the invention.

The primers, such as described in Tables 1 and 3, or including junctions of Table 2, or more generally including the flanking sequences of one of the 24 exons are also a part of the invention.

The kit for the detection of a predisposition to LGMD2 by nucleic acid amplification is also in the scope of the invention, such a kit comprises a least PCR primers selected from the group of :

- 10 a) in those described in table 1
- b) in those described in table 3
- c) those including the introns-exons junctions of Table 2.
- d) derived from primers defined in a),b) or c).

The nucleic acid sequence of claim 1 to 3 might be inserted in a viral or a retroviral vector, said vector being able to transfect a packaging cell line.

The packaging transfected cell line, might be used as a drug for gene therapy of LGMD2.

The treatment of LGMD2 disease by gene therapy is implemented by a pharmaceutical composition containing a component selected from the group of :

- 20 a) a nucleic acid sequence according to claims 1 to 4,
- b) a cell line according to claim 24,
- c) an aminoacid sequence according to claims 5 to 9.

REFERENCES

- 5 Arikawa, E., Hoffman, E. P., Kaido, M., Nonaka, I., Sugita, H. and Arahata, K. (1991). The frequency of patients with dystrophin abnormalities in a limb-girdle patient population. *Neurology* 41, 1491-1496.
- 10 Bashir, R., Strachan, T., Keers, S., Stephenson, A., Mahjneh, I., Marconi, G., Nashef, L. and Bushby, K. M. D. (1994). A gene for autosomal recessive limb-girdle muscular dystrophy maps to chromosome 2p. *Hum. Mol. Genet.* 3, 455-457.
- 15 Beckmann, J. S., Richard, I., Hillaire, D., Broux, O., Antignac, C., Bois, E., Cann, H., Cottingham, R. W., Jr., Feingold, N., Feingold, J., Kalil, J., Lathrop, G. M., Marcadet, A., Masset, M., Mignard, C., Passos-Bueno, M. R., Pellerain, N., Zatz, M., Dausset, J., Fardeau, M. and Cohen, D. (1991). A gene for limb-girdle muscular dystrophy maps to chromosome 15 by linkage. *C. R. Acad. Sci. Paris. III* 312, 141-148.
- 20 Birnstiel, M. L., Busslinger, M. and Sturb, K. (1985). Transcription termination and 3' processing: The end is in site! *Cell* 41, 349-359.
- 25 Blackwell, T. K. and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science* 250, 1104-1110.
- Bonilla, E., Samitt, C. E., Miranda, A. F., Hays, A. P., Salviati, G., DiMauro, S., Kunkel, L. M., Hoffman, E. P. and Rowland, L. P. (1988). Duchenne muscular dystrophy: deficiency of dystrophin at the muscle cell surface. *Cell* 54, 447-452.
- 30 Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* 212, 563-578.

- Bushby, K. M. D. (1994). Limb-girdle muscular dystrophy. In Diagnostic criteria for neuromuscular disorders, A. E. H. Emery, ed. (Baarn, The Netherlands: ENMC), pp 25-31.
- 5 Croall, D. E. and Demartino, G. N. (1991). Calcium-activated neutral protease (calpain) system: structure, function, and regulation. *Physiol. Rev.* **71**, 813-847.
- Dynan, W. S. and Tjian, R. (1983). The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* **35**, 79-87.
- 10 Emery, A. E. H. (1991). Population frequencies of inherited neuromuscular diseases - a world survey. *Neuromuscular Disorders* **1**, 19-29.
- Emori, Y., Ohno, S., Tobita, M. and Suzuki, K. (1986). Gene structure of calcium-dependent protease retains the ancestral organization of the calcium-binding protein gene. *FEBS lett.* **194**, 249-252.
- 15 Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246.
- 20 Fougerousse, F., Broux, O., Richard, I., Allamand V., Pereira de Souza, A., Bourg N., Brenguier L., Devaud C., Pasturaud P., Roudaut C., Chiannilkulchai N., Hillaire D., Bui H., Chumakov I., Weissenbach J., Cherif D., Cohen D. and J. S. Beckmann (1994). Mapping of a chromosome 15 region involved in Limb-Girdle Muscular Dystrophy. *Hum. Mol. Genet.* **3**, 285-293.
- 25 Goll, D. E., Thompson, V. F., Taylor, R. G. and Zalewska, T. (1992). Is Calpain activity regulated by membranes and autolysis or by calcium and calpastatin? *BioEssays* **14**, 549-556.
- 30 Gosset, L. A., Kelvin, D. J., Sternberg, E. A. and Olson, E. (1989). A new myocyte-specific enhancer-binding factor that recognizes a conserved element associated with multiple muscle-specific genes. *Mol. Cell. Biol.* **9**, 5022-5033.

- Hirai, S., Kawasaki, H., Yaniv, M. and Suzuki, K. (1991). Degradation of transcription factors, c-Jun and c-Fos, by calpain. *FEBS lett.* 1, 57-61.
- 5 Imajoh, S., Kawasaki, H. and Suzuki, K. (1986). Limited autolysis of calcium-activated neutral protease (CANP): reduction of the Ca^{2+} requirement is due to the NH₂-terminal processing of the large subunit. *J. Biochem.* 100, 633-642.
- Jackson, C. E. and Carey, J. H. (1961). Progressive muscular dystrophy:
10 autosomal recessive type. *Pediatrics* 77-84.
- Keen, J., Lester D., Inglehearn, C., Curtis, A. and Bhattacharya, S. (1991). Rapid detection of single base mismatches as heteroduplexes on Hydrolink gels. *Trends Genet.* 7, 5.
- 15 Kosak, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* 12, 857-872.
- Lovett, M., Kere, J. and Hinton, L. M. (1991). Direct selection: a method for the
20 isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. USA* 88, 9628-9632.
- Matsumara, K., Tomé F. M. S., Collin H., Azibi K., Chaouch M., Kaplan J-K., Fardeau M. and Campbell K., P. (1992). Deficiency of the 50K dystrophin-associated glycoprotein in severe childhood autosomal recessive muscular
25 dystrophy. *Nature* 359, 320-322.
- Matsumura, K., Nonaka, I. and Campbell, K. P. (1993). Abnormal expression of dystrophin-associated proteins in Fukuyama-type congenital muscular dystrophy.
30 *Lancet* 341, 521-522.

- Minty, A. and Kedes, L. (1986). Upstream regions of the human cardiac actin gene that modulate its transcription in muscle cells: presence of an evolutionarily conserved repeated motif. *Mol. Cell. Biol.* 6, 2125-2136.
- 5 Miyamoto, S., Maki, M., Schmitt, M. J., Hatanaka, M. and Verma, I. M. (1994). TNF- α - induced phosphorylation of IKB is a signal for its degradation but not dissociation from NF-KB. *Proc. Natl. Acad. Sci. USA* *in press*.
- Morton, N. E. and Chung, C. S. (1959). Formal genetics of muscular dystrophy.
10 *Am. J. Hum. Genet.* 11, 360-379.
- Murachi, T. (1989). Intracellular regulatory system involving calpain and calpastatin. *Biochemistry Int.* 18, 263-294.
- 15 Ohno, S., Emori, Y., Imajoh, S., Kawasaki, H., Kisaragi, M. and Suzuki, K. (1984). Evolutionary origin of a calcium-dependent protease by fusion of genes for a thiol protease and a calcium-binding protein? *Nature* 312, 566-570.
- Ohno, S., Minoshima, S., Kudoh, J., Fukuyama, R., Shimizu, Y., Ohmi-Imajoh, S.,
20 Shimizu, N., Suzuki, K. (1989). Four genes for the calpain family locate on four different chromosomes. *Cytogen. Cell Genet.* 51, 1054.
- Passos-Bueno, M.-R., Richard, I., Vainzof, M., Fougerousse, F., Weissenbach, J., Broux, O., Cohen, D., Akiyama, J., Marie, S. K. N., Carvalho, A. A.,
25 Guilherme, L., Kalil, J., Tsanaclis, A. M., Zatz, M. and Beckmann, J. S. (1993). Evidence of genetic heterogeneity in the autosomal recessive adult forms of limb-girdle muscular dystrophy following linkage analysis with 15q probes in Brazilian families. *J. Med. Genet.* 30, 385-387.
- 30 Richard, I., Broux, O., Chiannikulchai, N., Fougerousse, F., Allamand, V., Bourg, N., Brenguier, L., Devaud, C., Pasturaud, P., Roudaut, C., Lorenzo, F., Sebastiani-Kabatchis, C., Schultz, R. A., Polymeropoulos, M. H., Gyapay, G.,

Auffray, C. and Beckmann, J. (1994). Regional localization of human chromosome 15 loci. *Genomics* 23, 619-627.

5 Sambrook, J., Fritsh, E. F. and Maniatis, T. (1989). Molecular cloning: a laboratory manual. Cold spring Harbor Laboratory Press, Cold spring Harbor, USA.

Shapiro, M. and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene
10 expression. *Nucleic Acids Res.* 15, 7155-7174.

Sorimachi, H., Imajoh-Ohmi, S., Emori, Y., Kawasaki, H., Ohno, S., Minami, Y. and Suzuki K. (1989). Molecular cloning of a novel mammalian calcium-dependant protease distinct from both m- and mu- type. Specific expression of
15 the mRNA in skeletal muscle. *J. Biol. Chem.* 264, 20106-20111.

Sorimachi, H., Ishiura, S. and Suzuki, K. (1993a). A novel tissue-specific calpain species expressed predominantly in the stomach comprises two alternative splicing products with and without Ca^{2+} -binding domain. *J. Biol. Chem.* 268,
20 19476-19482.

Sorimachi, H., Toyama-Sorimachi, N., Saido, T. C., Kawasaki, H., Sugita, H., Miyasaka, M., Arahata, K., Ishiura, S. and Suzuki, K. (1993b). Muscle-specific calpain, p94, is degraded by autolysis immediately after translation, resulting in
25 disappearance from muscle. *J. Biol. Chem.* 268, 10593-10605.

Staden, R. (1982). An interactive graphic program for comparing and aligning nucleic acid and amino acid sequences. *Nucleic Acids Res.* 10, 2951-2961.

30 Suzuki, K. and Ohno, S. (1990). Calcium activated neutral protease. Structure-function relationship and functional implications. *Cell Struct. Funct.* 15, 1-6.

- Tagle, D. A., Swaroop, M., Lovett, M. and Collins, F. S. (1993). Magnetic bead capture of expressed sequences encoded within large genomic segments. *Nature* 361, 751-753.
- 5 Tomé, F. M. S., Evangelista T., Leclerc A., Sunada Y., Manole E., Estournet B., Barois A., Campbell K. P. and Fardeau M. (1994). Congenital muscular dystrophy with merosin deficiency. *C. R. Acad. Sci. Paris* 317, 351-357.
- 10 Uberbacher, E. C. and Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* 88, 11261-11265.
- Walton, J. N. and Nattrass, F. J. (1954). On the classification, natural history and traitment of the myopathies. *Brain* 77, 169-231.
- 15 Wang, K. W., Villalobo, A. and Roufogalis, B. D. (1989). Calmodulin-binding proteins as calpain substrates. *Biochem. J.* 262, 693-706.
- 20 Young, K., Foroud, T., Williams, P., Jackson, C. E., Beckmann, J. S., Cohen, D., Conneally, P. M., Tischfield, J. and Hodes, M. E. (1992). Confirmation of linkage of limb-girdle muscular dystrophy, type-2, to chromosome 15. *Genomics* 13, 1370-1371.

CLAIMS

1. A nucleic acid sequence comprising :
 - 1) the sequence represented in Figure 8; or
 - 2) the sequence represented in Figure 2; or
 - 5 3) a part of the sequence of Figure 2 with the proviso that it is able to code for a protein having a calcium dependant protease activity involved in a LGMD2 disease ; or
 - 4) a sequence derived from a sequence defined in 1), 2) or 3) by substitution, deletion or addition of one or more nucleotides with the proviso that
 - 10 said sequence still codes for said protease.
2. A nucleic acid sequence that is complementary to a nucleic acid sequence according to claim 1.
3. A nucleic acid sequence comprising in its structure a nucleotidic sequence according to claim 1 or 2, under the control of regulatory elements,
- 15 and involved in the expression of calpain activity in a LGMD2 disease.
4. A nucleic acid sequence encoding the aminoacid sequence represented in Figure 2.
5. An amino acid sequence which is coded by a nucleic acid sequence according to claims 1 to 4.
- 20 6. An amino acid sequence according to claim 5, characterized in that it is a calcium dependent protease enzyme belonging to the calpain family, involved in the etiology of LGMD2.
7. An aminoacid sequence according to claim 5 or 6 such as represented in Figure 2.
- 25 8. An amino acid sequence according to claims 5 to 7, characterized in that the amino acid sequence is modified by deletion, insertion and/or replacement of one or more amino acids with the proviso that such aminoacid sequence has the calpain activity involved in LGMD2 disease.
9. An amino acid sequence according to claim 6 to 8, characterized in that
- 30 LGMD2 is LGMD2A.
10. A host cell unable to express a calpain enzyme activity, characterized in that it is transformed or transfected with a nucleic acid sequence comprising all or part of the nucleic acid sequence according to any one of claims 1 to 4.

11. Use of a nucleic acid according to one of claims 1 to 4 or a host cell according to claim 10 in the manufacturing of a drug for the prevention or the treatment of an LGMD2 disease.

12. Use of an amino acid sequence according to claims 5 to 8 in the
5 manufacturing of a drug for the prevention or the treatment of an LGMD2 disease.

13. Use according to claims 10 or 11, characterized in that LGMD2 is LGMD2A.

14. Use of an amino acid sequence according to claims 5 to 9 for the
10 screening of the ligands of said amino acid sequence, said ligand being selected in a group consisting of substrate(s), co-factors or regulatory components.

15. Use of a nucleic acid sequence according to one of claims 1 to 4 in a screening method for the determination of the components which may act on the regulation of gene expression of calpain.

16. Use of an host cell according to claim 10 in a screening method for the
15 determination of components active on the expression of the calpain.

17. A method for detecting of a predisposition to a LGMD2 disease in a family or a human being, such method comprising the steps of :

- selecting one or more exons or their flanking sequences of the gene,
- 20 - selecting primers specific for these exons, or their flanking sequences, or an hybrid thereof,
- amplifying the nucleic acid sequences with these primers, the substrate for this amplification being the DNA of a human being; and
- comparing the amplified sequence to the corresponding sequence derived
25 from Figure 2 or Figure 8.

18. The method according to claim 17, characterized in that the primers are those selected from the group of :

- a) those described in Table 1;
- b) those described in Table 3; and
- 30 c) those including the introns-exons junctions of Table 2;
- d) those derived from the primers in a), b), or c).

19. The method according to claim 17 or 18, characterized in that LGMD2 is LGMD2A.

20. A kit for the detection of a predisposition to LGMD2 by nucleic and amplification characterized in that it comprises primers selected from the group of :

- a) those described in Table 1;
- 5 b) those described in Table 3; and
- c) those including the introns-exons junctions of Table 2;
- d) those derived from the primers in a), b) or c).

21. Packaging cell lines tranfected by a recombinant vector, characterized in that the vector contains a nucleic acid sequence as claimed in claims 1 to 4.

- 10 22. Use of a packaging cell line according to claim 21 as a drug for gene therapy of an LGMD2 disease.

23. Pharmaceutical composition for the treatment of an LGMD2 disease characterized in that in contains a component selected from the group of :

- a) a nucleic acid sequence according to claims 1 to 4,
- 15 b) a cell line according to claim 24,
- c) an aminoacid sequence according to claims 5 to 9.

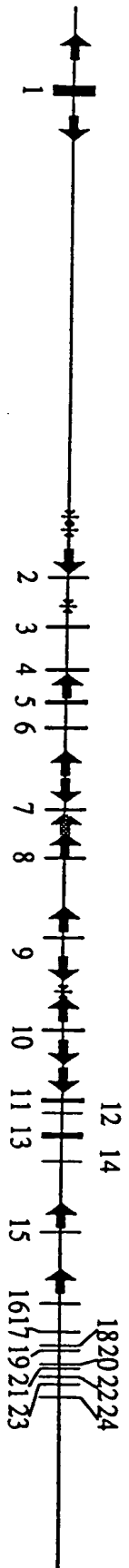
ABSTRACT

A nucleic acid sequence comprising :

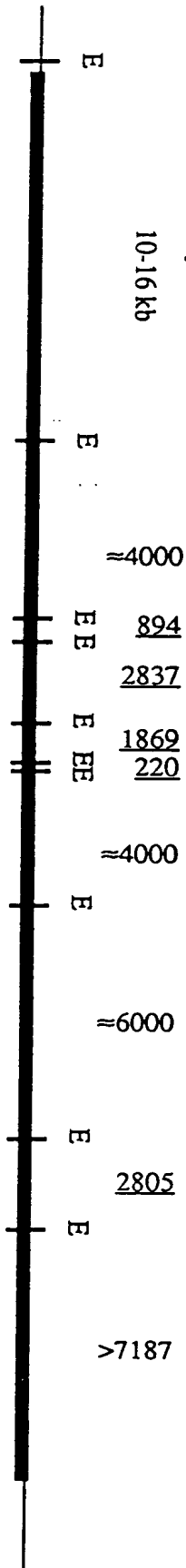
- 1) the sequence represented in Figure 8; or
- 5 2) the sequence represented in Figure 2; or
- 3) a part of the sequence of Figure 2 with the proviso that it is able to code for a protein having a calcium dependant protease activity involved in a LGMD2 disease ; or
- 10 4) a sequence derived from a sequence defined in 1), 2) or 3) by substitution, deletion or addition of one or more nucleotides with the proviso that said sequence still codes for said protease.

7
Figure 1:

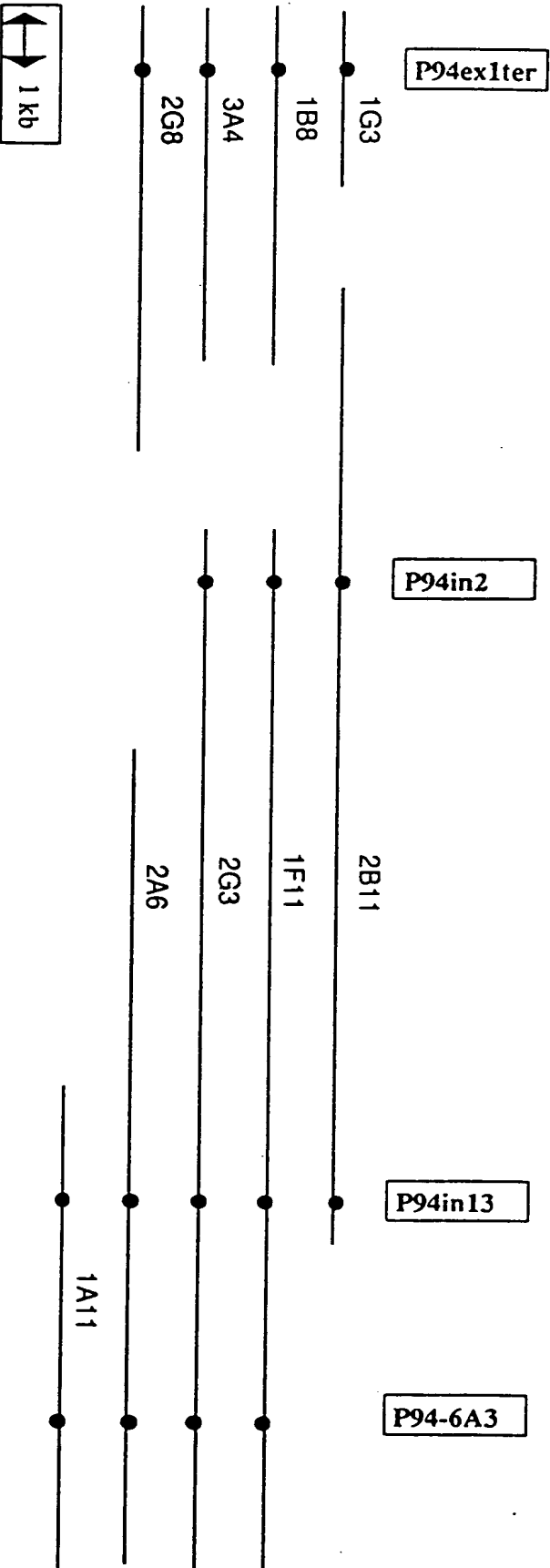
A) Genomic structure of the nCL1 gene



B) EcoRI restriction map



C) Cosmid map



A

[illegible]**B**[illegible]

C

[illegible]

Figure 3:

Human	1	MPTVISASVAPRTAAERSPGVPHPAOSKATEAGGNRPSGLYSALISRNFPILIGVKEKTEEQ	50	LHHKCKLEKKV	100	LVVDPEEPDETS	150	LFYSQKFP	200	IQFVWKRP
Rat	2G.....G.T.....H.G.....L.....								
Pig	3PT.....G.....G.T.....H.G.....L.....								
Cow	4PT.....G.....G.T.....H.G.....L.....								
Human	1	EICENPREIIDGANRTD	150	ICGEGDCMFLAAIACTLNQHLFRVIEBHD	200	OSFIENYAGIFHEQFWRYGEMVDVVIDDCEFTYNNQ	250	LVETKSNHRNEEWSALLE	300	
Rat	2G.....D.....L.....ER.....T.....D.....								
Pig	3G.....D.....L.....ER.....T.....D.....								
Cow	4G.....D.....L.....ER.....T.....D.....								
Human	1	KAYAKLHGSYEALKGGNTTEAMEDF	250	TGGVAEFFEIRDAESDMYKIMKKAIERGS	300	IMCCSIDDGTNMTYGTSPGLNMGELIARMVNMMDNSLLODS	350	SDLDPRGS	400	
Rat	2T.....K.....R.....								
Pig	3T.....K.....R.....								
Cow	4T.....K.....R.....								
Human	1	DERPTRILPVOYE	350	ETRMACGVIRGHAYSVT	400	GLDEVFPFKEKVKIVRLRNPWROVEWNGS	450	SDRWKDM	500	SFVDKDEKARLQHVTE
Rat	2V.....								
Pig	3V.....F.....								
Cow	4M.V.....F.....								
Human	1	ICNL	450	TADALQSDKLOTWTVSVNEG	500	FWRGCSAGGCRNFPD	550	TFWTFN	600	PCYRLKLEEDDDPDDSEVIC
Rat	2E.....								
Pig	3E.....								
Human	1	NKQH	550	LQKDFFLYNASKARSKTYIN	600	REVSQRFRLPEPSEYVIVPESTYEE	650	HOEGEFILR	700	FESEKRNLSSEVENTISVDRPVKKKKTKPLIFVSDRANSNKLGV
Rat	2R.....								
Pig	3R.....								
Human	1	QSEEGKGKTS	650	PDKOKSPOPOGSSDOESEEQQOFRNIFKQIAGDDME	700	LCADELKKV	750	LNTVVNKKHDKLTH	800	CGFTLESCRSRMIALMD
Rat	2D.....G.....GE.....R.....HT.....								
Pig	3D.....G.....GE.....R.....HT.....								
Human	1	NKIKAWQK	750	IFKHYDTQSG	800	TINSYEMENAVNDAG	850	FLHNNQLYDIIT	900	MYADKHMNID
Rat	2H.....								
Pig	3H.....								
Cow	4H.....								

Figure 4:

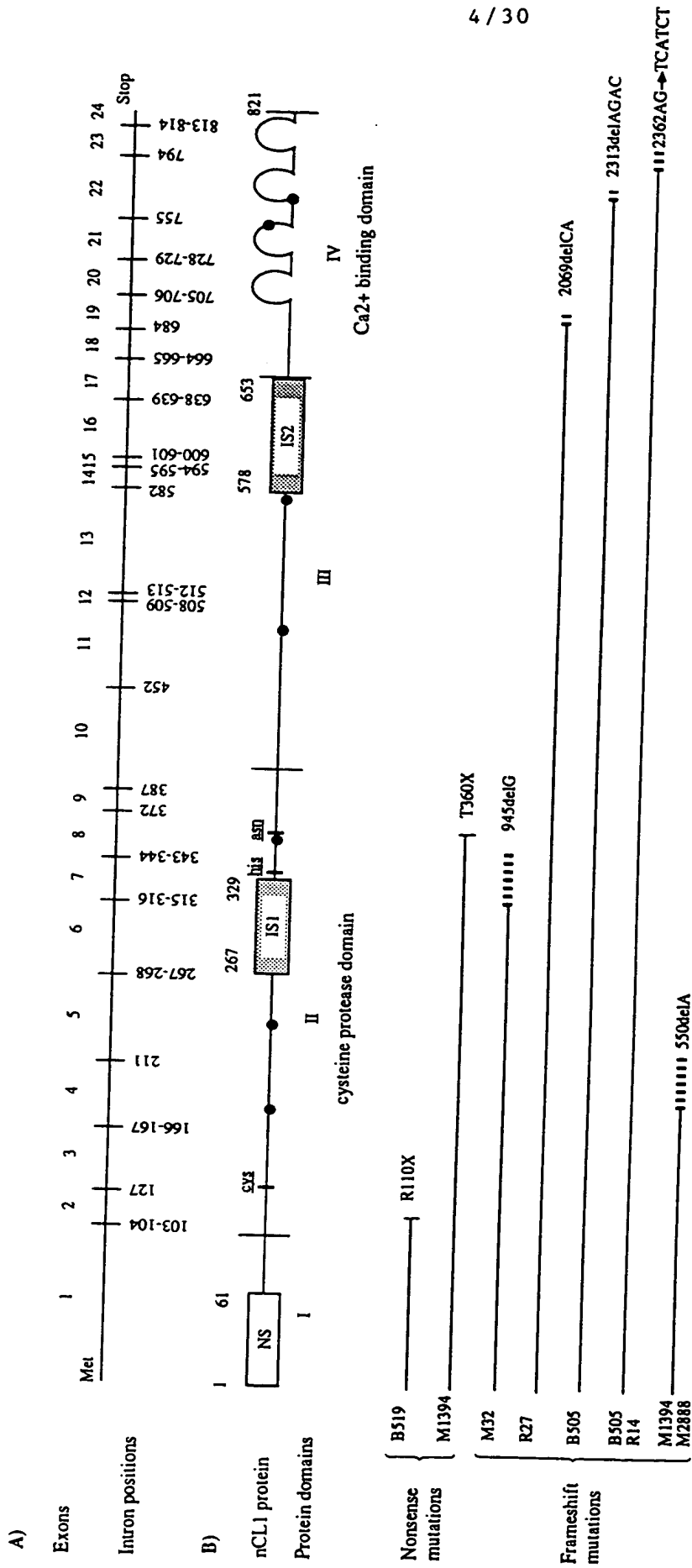


FIGURE 5

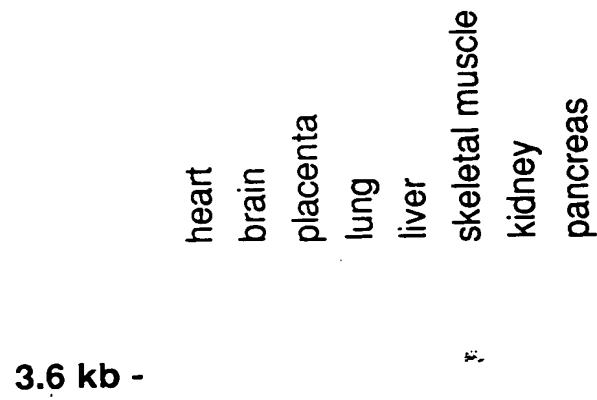


FIGURE 6

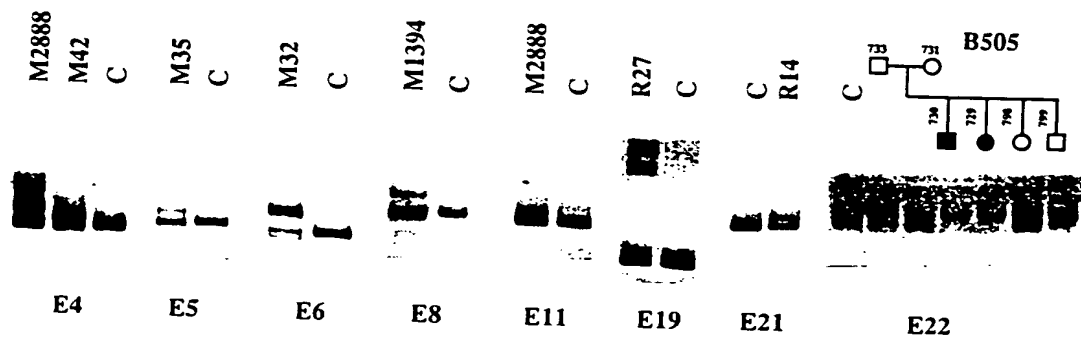
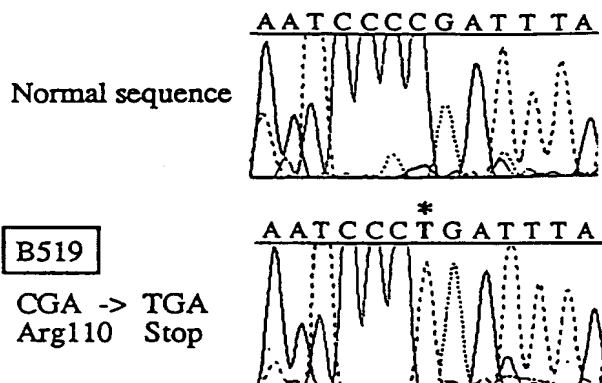


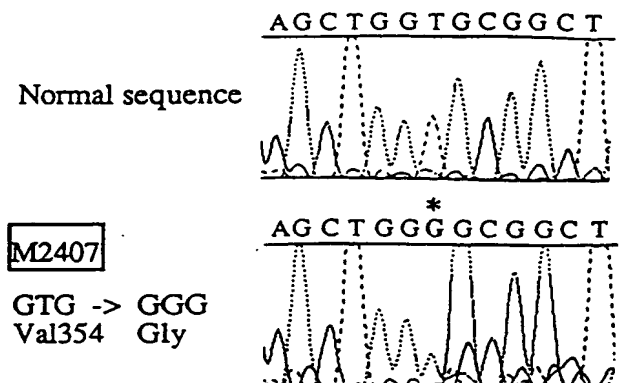
Figure 7:

7/30

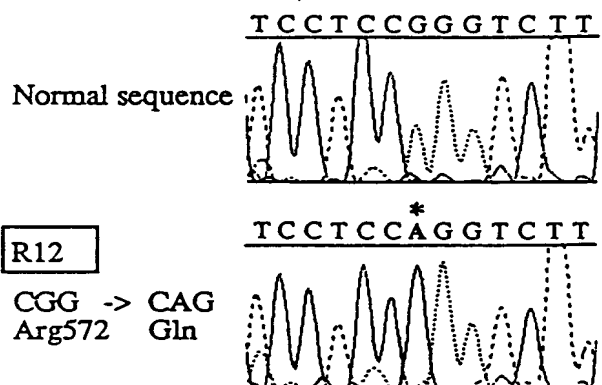
A) EXON 2



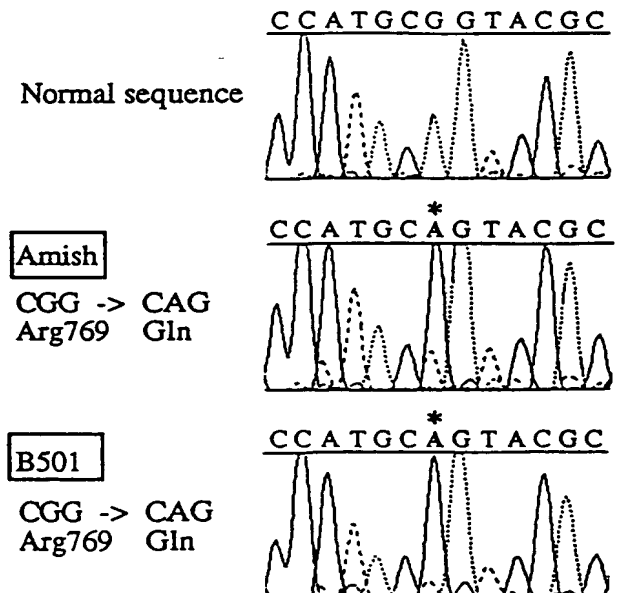
B) EXON 8



C) EXON 13



D) EXON 22



LISTE DE SEQUENCES

(1) INFORMATION GENERALE:

(i) DEPOSANT:

- (A) NOM: AFM
- (B) RUE: 13, place de Rungis
- (C) VILLE: PARIS
- (E) PAYS: FRANCE
- (F) CODE POSTAL: 75013
- (G) TELEPHONE: (1) 45 65 13 00

(ii) TITRE DE L' INVENTION: LGMD GENE

(iii) NOMBRE DE SEQUENCES: 4

(iv) FORME LISIBLE PAR ORDINATEUR:

- (A) TYPE DE SUPPORT: Floppy disk
- (B) ORDINATEUR: IBM PC compatible
- (C) SYSTEME D' EXPLOITATION: PC-DOS/MS-DOS
- (D) LOGICIEL: PatentIn Release #1.0, Version #1.25 (OEB)

(2) INFORMATION POUR LA SEQ ID NO: 1:

(i) CARACTERISTIQUES DE LA SEQUENCE:

- (A) LONGUEUR: 3018 paires de bases
- (B) TYPE: acide nucléique
- (C) NOMBRE DE BRINS: double
- (D) CONFIGURATION: linéaire

(ii) TYPE DE MOLECULE: ADN (génomique)

(xi) DESCRIPTION DE LA SEQUENCE: SEQ ID NO: 1:

TGATAGGTGC TTGTAACTG TGCTTAACGA AAACATACCG TGTGCTGTAG GGA	60
CTTGTTTATA TCAGTTAGCC TGGTTTCGCT AACAGTACAT CATTTTGCTT AAAGTCACAG	120
CTTACGAGAA CCTATCGATG ATGTTAAGTG AGGATTTTCT CTGCTCAGGT GCACTTTTTT	180
TTTTTTTTTAA GACGGAGTCT CTTTCTGTCA CCTGGGCTGG AGTGCAGTGG CGTGATCTGG	240
GTTACAACA ACCTCTGCCT CCTGGGTCA AGCAATTCTT CTGTCTCAGC CTCCCAAGTA	300
GCTGGGATTA CAGGCACCCG CCGCCACACC CGGCTTATTT TTGTATTTTT AGTAGAGACA	360
GGGTTTCACT ATTGTTGACC ATGCTGGTCT CGAACTCGTG ACCTCATGTG ATCCACCCGC	420
CTCGGCCTCC CAAAGTGCAG AGATTAGAGA CGTGAGCCAC ATGGCCCAGC AGGACCACTT	480

TTTAGCAGAT TCAGTCCCAG TGTTCAATTT GTGGATGGGG AGAGACAAGA GGTGCAAGGT	540
CAAGTGTGCA GGTAGAGACA GGGATTTTCT CAAATGAGGA CTCTGCTGAG TAGCATTTTC	600
CATGCAGACA TTTCCAATGA GCGCTGACCC AAGAACATTC TAAAAAGATA CCAAATCTAA	660
CATTGAATAA TGTTCTGATA TCCTAAAATT TTAGGACTAA AAATCATGTT CTCTAAAATT	720
CACAGAATAT TTTTGTAGAA TTCAGTACCT CCCGTTACAC CTAAGTAGCT TTTTGTGAAT	780
ATTGTTTTCC ATTCATTTGA TGGGCAGTAG TTGGGTGGTC TGTATAACTG CCTACTCAAT	840
AACATGTCAG CAGTTCTCAG CTTCTTTCCA GTGTTACCT TACTCAGATA CTCCCTTTTC	900
ATTTTCTGTC AACACCAGCA CTTCAATGTC ACAGAAATGT CCCTAGCCAG GTTCTCTCTC	960
TACCATGCAG TCTCTCTTGC TCTCATACTC ACAGTGTTC TTCACATCTA TTTTITAGTTT	1020
TCCTGGCTCA AGCATCTTCA GGCCACTGAA ACACAACCCT CACTCTCTTT CTCTCTCCCT	1080
CTGGCATGCA TGCTGCTGGT AGGAGACCCC CAAGTCAACA TTGCTTCAGA AATCCTTTAG	1140
CACTCATTTT TCAGGAGAAC TTATGGCTTC AGAATCACAG CTCGGTTTTT AAGATGGACA	1200
TAACCTGTCC GACCTTCTGA TGGGCTTTCA ACTTTGAACT GGATGTGGAC ACTTTTCTCT	1260
CAGATGACAG AATTACTCCA ACTTCCCTT TGCAGTTGCT TCCTTTCCTT GAAGGTAGCT	1320
GTATCTTATT TTCTTTAAAA AGCTTTTTTCT TCCAAAGCCA CTTGCCATGC CGACCGTCAT	1380
TAGCGCATCT GTGGCTCCAA GGACAGCGGC TGAGCCCCGG TCCCCAGGGC CAGTTCCTCA	1440
CCCGGCCCAG AGCAAGGCCA CTGAGGCTGG GGGTGGAAAC CCAAGTGGCA TCTATTGAGC	1500
CATCATCAGC CGCAATTTTC CTATTATCGG AGTGAAAGAG AAGACATTCG AGCAACTTCA	1560
CAAGAAATGT CTAGAAAAGA AAGTTCTTTA TGTGGACCCT GAGTTCAC CCGATGAGAC	1620
CTCTCTCTTT TATAGCCAGA AGTTCCCAT CCAGTTCGTC TGCAAGAGAC TCCGGTGAGT	1680
AGCTTCCTGC TTGCTGGCTG GGTTCCTCCC CCACGGAGGA GTCCTCTCAC TCAGCACCTC	1740
CGGCAGCTCA GCTGTGCACA TGGGCACTGG GGGAAGGATC CTGGCAGCAG CTCTGCTGGG	1800
CTCTGTCTTT AAGTGTGAAG CAGGGAGGAG AGGAACAGGT CTCAGATATT TCACCAAATC	1860
TCAGCAAAT CCAGAGGGAG AGCGCAGGAG GTGGGGTGAT TCTTATGCTC TGGCTCTTTC	1920
TCTCTGAAAA AAAAAAAAAA ATCTTGCTTT TTATAAAAGT GGGTGAACT CAGTTTAATT	1980
CATCCTGTAA AAATAAATAT TCCTTTCTCA GAACAAATTC CAGACAGCCC AGATGTACCT	2040
GTTCGTTTTA ATATTATTCA TCTTGGTAAG ATTATTTTCA TTTCTCTGGC TAAAATCATG	2100

FIG 8A/2

ATGTTATTCT TCTTTAATTT ACCAATGGCC ATTCTTTCTG AAACACAGAA ACCCTAGAAA	2160
GAGAAGAGTC ATAGGCAAGG AATTTTTTTC ATGCATAAAA TGTTGGGGTT AAAGAGAGAG	2220
AGACCTAGCA ATCGCTTTGG TCCACCTACC TCACCTCATA AGTGAGGAGT CAAGGCACAC	2280
TAGAGTGAAA TATATCTAGT GGGCACATGA CAGAGCCCGG ATTAAAACTT TGTTTTAGGA	2340
AACTCTCCCA GCCTCTGGGT TTCATTTACA GTGATCGCCA GGAGGGAAAT CACATTCCCC	2400
TGGCTCACCT CTCTGATCAT CCCTCCAGTG TGA CTCTTGT TCTTAATTCG AGAAATATTT	2460
ATTGAGCATC TACTAGTGCC AGCACTGGGC AAGCAACTGG GGGGACAGCA GTGAGTAAGA	2520
AAGACCAAAA TTCCAGCTGT CTTGGAACCT AGGGTCCTGA AGGGAAGATG GGCATTGAAC	2580
AAGAGTGACA TTGTCAGGAG ACGATGTTCT GGGTGCCACA GGATCATGTG GCAAGGAGAG	2640
CTAACCTGGT CCAGGGAGAC AAACCCTCTC TGAGGAAATG ATGACAAGCT GAGACCCAAT	2700
ACTATTGATT AGCCATGGTT TTCTTTAACC TAAGGTGGGC CAGGCATGGT GGCTCATGCC	2760
TATAAACCCA GCATTTTGGA AGGCCCAGGC TGGAGGATTG CTTGAGCCCA AGAGTTAGAG	2820
ACCAGCCTGG GCAACAGGGT GAAAACCTAT CTCTTTTGTA CTAAAAATTC AAAAAATTAT	2880
CCAGGCATGG TGGCAGATGC CTGTGGTCCT AGCTACTCAG AGGCTGAGGT GGAAGATCA	2940
CTTGAACTCG GGGAGTTTGA GGCAGCAGTG AGCCGAGATC ATGCCACTGC ACTCCAGGCT	3000
GGGTGACAGG AGTGAGAC	3018

(2) INFORMATION POUR LA SEQ ID NO: 2:

(i) CARACTERISTIQUES DE LA SEQUENCE:

- (A) LONGUEUR: 11451 paires de bases
- (B) TYPE: acide nucléique
- (C) NOMBRE DE BRINS: double
- (D) CONFIGURATION: linéaire

(ii) TYPE DE MOLECULE: ADN (génomique)

(xi) DESCRIPTION DE LA SEQUENCE: SEQ ID NO: 2:

GATCCACCCG CCTTGGCCTC CCAAAGTGCT GAGATTACAG GTGTGAGCCA CCACGCCCAG	60
CCGACACTGC CCTAACTCTC AAGTTGCATC CTTACTCGAA TAGTATGACA GTGTGGGAAG	120
CAGCATGGGA CAATGTAAAA AGGAGGCATG TTTCTGGCTT CTGCTACTTA CTAGCTGTGT	180
GTCTTTGCAC GAGTTTCTTA ACCTCTCTGG GCCTCAGTTT CCTTATCTGA AAAATAACAA	240
TGATAGTATT CCCTTCACAG GGCCAAATGG AATACTATCA GGAACACTAC ATAATGGAAC	300
TCAATAAATA ATAGCTACTG CGGCCGGGCG CGGTGGCTCA CATCTGTAAT CCCAGCACTT	360
TGGGAGGCCG AGGCGGGTGG ATCACAAGGT CAAGAGATGG AGACCATCCT GGCCAACATG	420
GTGAAACCGT ATCTCTACTA AAGATACAAA AATTAGCTGG GCATGGTGGC GCATGCCTAT	480
AGTCCAGCT ACTCGAGAGG CTGAGGCAGG AGAATCACTT GAACCCCGGA GGCAGAGTTT	540
TCAGTGAGCC AAGATTGCAC CAGTGCCTG CAGCCTGGCG ACAGAGTGAG ACTCCGTCTC	600
AAAAAATAC CTATCTATCT ATCTGTCTAT CTACTGTTAT TCTTACCTGG TCATTTCTT	660
TTTGTTTCAC AGGAAATTG CGAGAATCCC CGATTTATCA TTGATGGAGC CAACAGAACT	720
GACATCTGTC AAGGAGAGCT AGGTAGGAAA GTGCCTCAGG TCAGATCCTG CCAGATGATC	780
AAGGGGTGAT TACAAGGTGT GATCCCTTC CAGGAGGTAA AGGGACAATC TGTGCTTGCT	840
TCCAGTAACT TTTTGAAGA TTTTTTATAA CAGTTGCTTT ATGGTCGTTT ATCTACATGC	900
TGGCGATTGC TTCATTTCTT CCTACATGCC TCTTTAGCAC TCTGCCATGC ATCACAGGGG	960
GTATCTGCAT CCTGTGGCCT CCTCTCCAGT ATCTCAAGGA CACTTACATA CCCCCTCAG	1020
CATGACAAAA GCCCTGCTTT TCACTGTATC GTCTTTCTTG GAAGACAGCT CTGTGACTGT	1080
GCACCAAGCA TGCCCTTGG GCATGGAGAT TCTAGATACA CACACAAAAG GCATCGCCAA	1140
GGAAAGCACT TGTAAGTGA ACCCTTGTT TAAATTGGCC CAGCATAGCT CCATCTTTAA	1200

AAGAGTCTTT CCACAAAGAT GGCATCCGCC ATGTGGATGA GCATCCAATT TTCTCTTTGA	1260
TTGGTTAGCT TGA CTGCTCC ATCTGATCTT CCTCTCTCTC GACCTCTTGT TCAGAAAGTA	1320
TTGTCTTTGG TGTGGACTAT AAGCAAGCTC TGTGAAGTAA AATTGGAGAG AACACCAACA	1380
GAAACAATTT AAATTTGAGG AAAAGGGGGC ACCTAAGACC AAAGGAATTT GGCTTATTTT	1440
ATTCCAGAAG GGGAGGCTGA GAATAAATCA GATGAATATC TGGGTTCCCTG CACCTGAGGG	1500
AAGGCTTCCT GCAGAGCCCT GGGCATAATA ATCTGGGACC TTCAAACCAA TAACCTCTTT	1560
TCCAAGGAAA GACTGGCTGC TTCCAAGGAG GGTAGGGGAG AGTCGGGCTG CAGGCAGCTC	1620
TCAAGTCTCC CCTTGACAC TCTCAGGTTG GCATTTTCAC TTTAACCCAT CCTCCCTTAA	1680
GAAGGCAGTT CTTTGTGACC AGGGTACACC CCCTATTATA TATATATATA CACACACAGA	1740
GAGAGAGAGA GAGAGAGAGA GAGAGAAAGA GAGCAAAGTG TTACCTCCAA CTACATACAG	1800
TACTCTGTCA GAAAAGAGGT TCAGAGAATA AGAAAACGTC CCGAGCTCAT TCCGTTGCCA	1860
GCAATGTCTT ACTGCCCCCT ATAGACGGGT TCCAGGGCAG CTGCCTACCT GGCCTTCCTT	1920
CCAATACAAA TCATCTTGGT GGATGGTTCT CTGAGGCTCA GTCTTCGCTG AAGTCAGAAG	1980
AGGAATTGGA CTCACATTGC AAAGGCACAG GGCAGGGCAG ATTTCTTACA GGTGTTAGGA	2040
AGAACAACCC AGTTATGATC ACCTACTGCT CTGTCTCCAT TGAGGCCTAA AAAGGAAGTG	2100
AGTTTATACT GCAGTTGGAG GAACTGCCTG CAGCCTTGAG GAAAATGTCT AGTCACAAGG	2160
GAGTAAGTTA CCTGTTGATC ATATTGTCAA GGAATTCCTG TCCAATTCTC CTTCCCTGGG	2220
TTGACACCTC TGTAAGGTCA GATCTGGAAG TAGGAGAGTG GGCACCAAGG GAGTCCCCGT	2280
TCAGGGAAGT GGAGTGGCTG GCTGGGATTG GGGCTTTTTT TCCCAGGAG GAGCAGGAGT	2340
GCTCAGGATC TGTGCCCTGT GTCTGCCTGC AGGGGACTGC TGGTTTCTCG CAGCCATTGC	2400
CTGCCTGACC CTGAACCAGC ACCTTCTTTT CCGAGTCATA CCCCATGATC AAAGTTTCAT	2460
CGAAAACCTAC GCAGGGATCT TCCACTTCCA GGTGAGGTAA TGAGAGTCTA GTTAAGAGGG	2520
CCAGCGGCAG GCCACCCACC GCTGGTCTCC TGGCCTTGAC TTCCCAGAAG CTGGAGGAAA	2580
CTTCCCACCC ATCTACCCGC AGCGGCAACA GTCGGCATGG ACCCCCTTAA GGCTTCAAGC	2640
CTGGGAGGAA GCAGTTGCTT ATCTCTGGCT CCCTAATCCC TCCCCACCA CCTTCCACTA	2700
TGTCCCAGAA AGACAGGAAG ACATCCTGTT TACTGTGGGT CTATTTTGT CTTTGCAGCT	2760
GTCTGGCTGC TTTTATTGCC TGCAGCCCTT CTCAAGTAGG TCCCTAAGAT ATTAGCACTG	2820

FIG. 8B/2

TGACACCACA GGACCCTTCA GGTGTACAG GAACCCCTGT CCAGGGCTCC TGTATACTTC	2880
TTCCTCTCTA AGGCATGGCG GTACCAAGGC TATCACTCCT CTCTTCCAAG CCCTGGAAGA	2940
AGAGTCTGCT TAACCTGGGG ATCAGGCTTC TTGTTTGCCC TAGAACTGAA TCTGATGGTT	3000
CTAGAATCCA TCCAGCTACT GGAAATTTTC TGGGTCCCAG TCACCTTGGC ATAGAGCTGG	3060
TGCTAGAGCA GAACCAAACCT GAATTCTACC TGTGAGGGTC TCGTAGCTTC CGGGATGCTG	3120
GGGAGTCAGC CTGTCTCCAG CTTCAAAGGC TCCCTCATGT CCCAGGATGA CCCACATTAT	3180
CAGTTCTTGC TCCCCGGGTC TTGCACCTCA GCACGGAAGG CCTCAGAAAA GGTCTGTCTC	3240
CAGGCTCAGA CTCCCCCTCC TGCCGCCTTG GGAACATGGC ATATTTAAAG GGTCTCAGAT	3300
CTAAAGGGCC TTACATACAA ATATCAGATA GATTTCTGTT CTCATTTCAG TGAGGGAGAA	3360
AGTGCCATTG AAAAGGAGAC TAAACCACAT TTGGCCCTTT TCAGTTCAAA CTGATTTCATT	3420
CAAAAAAGAG CGACATCCAA ACTTGAAATG ATTGAACAAT GTTCCTGCTA CAGCTAGAAT	3480
AGATTCTGGG TCACTTTGTT CCTCCGTTTC AATCCTTGTT CTTCACTTTG GCATCAAGAA	3540
ATACCTAAAT CAGCACAGTG CCTTCACTGC ATAGTTCCCA ATCCTGGCCA CATTGAATCA	3600
GCTGGGGGCA CCTGAGAGTG CTGACACCCA GGCCCTGCCC CAGACCTGCT GAGCAGGAGA	3660
ATGAAAATCT TACATCCTAA GACACTCATG GAGCACCTAC TCTACCCATT ACTGGGCTGG	3720
ACTCTGTGGA AGACATGAAG TATATGTAAC TCACTTCCAG CTCTCAAAAA GCACCCAGTC	3780
CAGTTAGAGA CAGATTTACA CACCCCAAAC ACAAATAGG ATGAACAGGC ACCCAGATGC	3840
AGAGTCCAGG AAATGATGCT GCTTTGGGAT TCAAGAACCC CCTGAGGAAT GTGGAGGAAG	3900
GACACATTC CTAAACAGTAA TTTGAGTATG TGAATCTGTG CGTGACGCTT CTGTGCAGTT	3960
CTGGCGCTAT GGAGAGTGGG TGGACGTGGT TATAGATGAC TGCCTGCCAA CGTACAACAA	4020
TCAACTGGTT TTCACCAAGT CCAACCACCG CAATGAGTTC TGGAGTGCTC TGCTGGAGAA	4080
GGCTTATGCT AAGTAAGCAA CACTTTAGAA TGTGAGGTGG GGCTAGAGGT GAGAAAGTGG	4140
GTTGCAAAAT CCAGCCGAGA CCTCACTCAC AGGAAGAGGC ATGTGCCTCT ATACGTGCAT	4200
ATGTGTGGGC ATGCAAGTCC AACTGTGACC CAAAGTTAGA GATCAGTTCC AGGCAACAAC	4260
AGCTCTAACT AAAAACATTA AATTTAAGAG TAGAAATGAA GATTTCATA GAAGACCTTT	4320
AGCTTTAGCT CACCATAGCG AGTTCTTTCA TTGCACCTCC ATGGTGGCAT TGCAAGTCTT	4380
GGGATCAGAG CATTGTCCCA GGTCTCGAT TGGCTCAACC TCATGTGCTT ATAGAAGATT	4440

TATAAAGACA TGTTGTCTCT CAACTTAAAA GCTCCACCCC AGATGATAAT AATGGATTTT	4500
CAAATTTTGG AACAAAGGTCA CTCTGTAATG CAGGCTGGAG TGCAGTGGTG CAGTCACGGA	4560
TCACTGTAGA TTGACCTCCT GGGTTCAAGG TGCTCCTCCC ACCTCAGCCT CCCAAGTAGC	4620
TGGGACTACA TGCGGGCATC ACCATGGCCC TTTTATTTTT GTATTTTTTT GTAGAGCGGG	4680
GTTTTCCCAT GTTGACCCAG ACTGTTCTCG AACTCTTGGG CTCATACAAT CCACCAGCCT	4740
TGCCCTCCCG AAGCGCTGGG ATTGCCGGTG TGAGCCACCA CACCGGCAGC TGCTAATGGC	4800
TTTAATGCAG CCCTTCCTCA ACGTTCAGGA TGTAGTGGA AGAGCTCTCA GGAAGTGGGG	4860
ATAGCTGGGT TTCAATCCCA GTGCTTCTGG CTCTCTGTGG TCTTGGGTGG GTCACCTAGC	4920
CTCTTGAGCT CAGTTTCTTC ATTATGAAGA AAGGGAATCA TTGTTTCCAT CCCATGAGCT	4980
CATAGGGTTA ATGTGGAATT GATGAAAGAA CATCACAGCA TCCAAGAGGT AAAGTTCTGG	5040
TGGCAGTGGT ACCTGGGTTT TGTTCCCTGG AACTCTGTGA CCCCAAATTG GTCTTCATCC	5100
TCTCTCTAAG GCTCCATGGT TCCTACGAAG CTCTGAAAGG TGGGAACACC ACAGAGGCCA	5160
TGGAGGACTT CACAGGAGGG GTGGCAGAGT TTTTGTAGAT CAGGGATGCT CCTAGTGACA	5220
TGTACAAGAT CATGAAGAAA GCCATCGAGA GAGGCTCCCT CATGGGCTGC TCCATTGATG	5280
TAAGTCTGGG GTGTGGGGCA CAGGGTGGGG AGCTCCAAGT GTCAGGAAGC CTTTTACCCA	5340
ATGAAGGGCA GCATAGAGCT TTTGTGTGGG ACAGAGCGAA TGTTTGTGT GAGGAAGCAG	5400
GAACTGGCTC TCAACTTTGA GGA CTGGGAA TTTCTCAAGG GAGAACAGTT CTTCCGGATT	5460
TTCAATAAAG AACTGGTCA AGGACATTTC AAGCCCTGGA ATGTCAGTGG AAATCAGTCC	5520
AGAGGCCTGT GTCAGTGGAG GCCTCCCTTG CTGGTGCTCC TCAGTCTCAG CACGCTCCCA	5580
TTAAGCTGGC CACGTACTTG GCTGTGGACC TGAGCCCACC ATTTCCCTAA GAAAGCCTCC	5640
CAGTCACTGG GCTTTCACCA CACCTCCCCG CTTGAGACGT GGGCTTTGTG TTGTTACCTG	5700
GGAGAAGCTA AGCCTGCAGC ACCTTTCAGT GCAAAGAAAT GCTGTGAACT GAGACAGGAG	5760
CCAAGGGTAG GGAGATGGCC GCCCATGGCC AGGCCTCCTT CAGGGGGCAT GCCTTCCCTG	5820
AGGGCTGCTC AGTATATTGA TATGATAATC TTAGTGTTT CCATTGGGGA GGATGGGGCT	5880
GAAGCTGAAT TCCTGCCCCCT TCTTCTCCCA ACACGCCCAA TGGACAGCTT GGAAGGTCAG	5940
TTAGCACACA ACACCATGGA TGAACTTTTT TTCTGTATCA CTTTCTCCG TCTTCTCTCC	6000
ATTCTGTCTC TGTTGATCTC TCCTCTCTCC CTTTGTCTGT CCCATCTCTT TCTCCTCTCT	6060

FIG. 8B/4

CCTTCCCTTT	CCACCCTTCT	GTGTTTGTTT	TCTCCCTCCC	CTGTGTTGTT	CCCTACATTC	6120
TCCATCGGGC	CTCAGGATGG	CACGAACATG	ACCTATGGAA	CCTCTCCTTC	TGGTCTGAAC	6180
ATGGGGGAGT	TGATTGCACG	GATGGTAAGG	AATATGGATA	ACTCACTGCT	CCAGGACTCA	6240
GACCTCGACC	CCAGAGGCTC	AGATGAAAGA	CCGACCCGGG	TGTGTACACC	TCCGATTATC	6300
AGAACTGACC	ATCCCTCCAA	CCCACATGAC	CCCGCCCTAT	TAGTGTGAGA	CTCCCCTCAG	6360
CAGCCAGGGC	CTTACCCACA	CACCCCGACC	TGGCACCTCC	CAAGGGTCTG	GGTTGAAATA	6420
ACTTGCTCAG	CCAAGGCTCC	TGAAGAGGGT	GCAAGAACCA	GGATTTTGGA	GGGAATCTCT	6480
GCTGGAGTTT	CTGCATATTC	CATGGTCCAG	GCAGTTCCTC	TCATAACGAA	CTATCAGACA	6540
GAAATACTTG	TAAAGATACT	TCATTTATTT	TGAAATATTT	TTCTCTTCT	AATGTATTCA	6600
TTTATTCATT	CAACACTTAT	TTTGAGCTC	CTACTATGTT	CCAGGCACTC	CTCTAGCAAA	6660
CAAAGCAAAT	TCTCTCCTCT	TTTTCAATAT	TTGTGGAAAA	AGCAAGGTCT	CCCTCTTGTA	6720
GAGTTTATAT	TCTAGTATTT	TCATAAGTTA	TACCTGCTCA	CTGGAGAATA	CTGAGCCATA	6780
CAGAAAAACA	CAGAGGAAAA	TTTCACTTAT	ATTTTTCCCC	ATGTAAAGAT	AACCACTCTT	6840
AACATCTAGT	ATATGTTCTT	CCAGGATTTT	TCTATGCACA	CACTGAATCT	GTATTTTTAT	6900
TTTTAAAATG	TTATCATATT	GTATGTACCT	CTTTGCAGCC	TGCTTTTTTC	AGTTAGTTTT	6960
TTTGGTTTTT	TGTTTTTTTT	TTTTTTTTTG	AAACCAAGTC	TTGCTCTATT	CCCTAGGCTG	7020
GAGCACAGTT	GTTGCCATCT	CGGCTCACTG	CAACCTCTGC	CTCCAAAGTT	AAACTAATTC	7080
TCCTGCCTCA	GCCTCCCGAC	ATAGCTGGGA	TTACAGGCAC	ACACCACCAC	ACATGGCTAA	7140
TTTTTGATTT	TTTTAGTAGA	GACGGGGTTT	CACCATGTTG	GCTGGAATGG	TCTTGAAGTC	7200
CTGACCTCAA	GTGATCCACC	TGCCTCAGCC	TCCCAAAGTG	CTGGGATTAC	AAGTGTAAGC	7260
CACCACACCC	GGCCTAGTTT	GATATTCTTA	ATGTGCCCAA	AGTATTCTCC	TGTAACATTT	7320
TTTAATAGCT	ACACAATATT	CAAAACACACA	GATATGTTAT	AATTTATTTA	CCCAATACCC	7380
TATTATTGGA	AAGTTGAGTT	CTTTTTTTTC	TTTGTTTTGT	TTTGTTTTGC	TACTATTCTA	7440
AAATGCTATA	ACGAACATCC	CAATAGATAC	ATCTTTGTAT	ACATCCATGG	TGACTTCCAT	7500
AGGACAGATT	CCCAGCAGTA	GAATTGCTGG	GTTGAATGAT	ATGCTTAGGG	TAATGACAGA	7560
AGAGTCATTT	CAAGCAGCTT	CCTAGGGTCT	TAGAACTTAA	GGATTAATGA	GTCTTCCCGC	7620
CCCCTCCCAG	TCTATTCAGC	ATGATCTGGA	TCATGAGGAC	TGAGATCTGG	AAGAGACTGA	7680

GATCTGGGAG AGGCTGAGAT ACCAAAAGCC CTGGCTCCAC CCATACCCCT CGCCCTGAAA	7740
ACAGCTCTAG GAATTCCGCG GCCTAGCAAG GCTCCGGGAA GCTCCTTTA AAGCTGTGAC	7800
GTTAGTAGGC ACATGGACCA TAGAGACCTA TCCAGGGCTC ATGGGACTTT AGTGATCCTG	7860
CCCTTCTCCC AAGGATCCCC CATGGCTGCA ACTTGAAAT TTCTGCAAAT GGAAGAGCTA	7920
CTCCTTAGGC ACGGTCATGT CTGAGCAGGG ATCTCCTCGG GCTTTCTTAG AATTCTCTCC	7980
CTGGGCACTG GGA CTCTTGA TTTCTTGAAT ATTATGTTCC AGGTGGGTGT GGAGGAGGTG	8040
AGGGGATGTA AAGAAGGCTA GACTTGGCCA GGGCAGTGG CTCATGCCTG TAATCCCAGC	8100
ACTTTGGGAG GCTGAGGCGG GTGGATCACC TGAGGTCAGG AGTTCGAGAC CAGCCTGGCT	8160
AACATGGTGA AACCCCGTTT CTAATAAAAA TACAAAAAAT TAGCTGAGCA TGGTGGCAGC	8220
TGCCTGTAAT CCCAGCTACT CGGGAGGCTG AGGCAGGAGT ATCGCTGGAA CACGGGAGGC	8280
AGAGATTGCA GTGACCCGAG ATCGCGCCAC TGCACTCCAG CCTGGGCGAC ACAGCAAGAC	8340
TCTGTCTCAA AAAACAAAAA AGAAAGAAAA AAAGGAAAAAG CTAAGACTTA CATGTGTCAC	8400
TTAACCCCTT TTCTCAAACC TCTTTCTCTT CCAGGAATAG TCAACCCCTG GATGGCTTCA	8460
GGGGAAGGGG GATCCTGAAG CCCAGGGCAG CCTCCAATC TACCCCTTCC TCCTTTGAAG	8520
GATACTAAGG GGTCCAGAAA GGAGGGGCAG GACACTGTTA CCCACCCAC ATCCAGCAT	8580
CCACATTGCT CTCTGATGGT CAGGACAGAG CTTTCTCAGG GAGACCAGCC TGTCTGGAGC	8640
TGTGTCTCTT GGCACCTTA AAGGGCCACT GAAGGTCCGT TCGTGGTCGT GAGGCACACT	8700
TTCAGGGAGC AGAGTGGTCT GTGTCTTCAC AGAGCCCGGA AAATGAACTA GTATGAACTT	8760
TGCCTCCAAG CAGCAGAACT TCTGTTCCCC CGCCCTAAT GGGTTCTCTG GTTACTGCTC	8820
TACAGACAAT CATTCCGTT CAGTATGAGA CAAGAATGGC CTGCGGGCTG GTCAGAGGTC	8880
ACGCCTACTC TGTCACGGGG CTGGATGAGG TAAGCCTGGT GGGGCTTGGT GGGGCAAGGG	8940
CACCCTCCTG GGTAAACCTC ATGAAGTCAG GACTTAGCTG TTGGGGCCCC TGCCCTGTCT	9000
GCAGAGCTTG CCTCCAATCA GGACATTCAG TTCAAGGTCC AAGCCACGCC TGGGAGCAGA	9060
GGGGCCTGTG AAAGTGGTAG AGGTGGATCC TGCCACAGTT GGTGCACAGT TTATCTTTGC	9120
TTTTCGTGCT AAAGATGGCA ATTTTCCAA CATTTCCAAT GAACAAATTG AAATATCACT	9180
TAACTTTGCT TTTACAAACT TGGTTTCATG TGTTCTTGAG CTTCTGTTC TCTCGTCTC	9240
AGATAGCTAC AGTTGTCTCT GGGTAGCCAC GGGGACTGGT TCCAGAAGCC CCAACAGTAA	9300

CAAAATCTGC	AGATGCTCAA	GTCCCTTCTG	TAAAATGGAG	TAGTATTTGC	ATATAACCTA	9360
TGCACATCCT	CCCATATACT	TTAAGTCATC	TCTGGATTAC	TTACGATACC	TAACACAATG	9420
GAAATGCTAT	GTAAATAGTT	ATTGCACTGC	ATTGGGTTTT	TTTGGTATTA	TTTTCTGTTG	9480
TTGTATTATT	ATTTTTTCTT	TTTTTGAATA	TTTTTGATCC	ACAATTGGTT	ATATGCCAAA	9540
GCCATGGATA	CGAGAGGCTG	ACTGTTCTGT	TTTGCTCCTT	CTGGGACTTC	TGGGTTTTTC	9600
TGGACCATGT	CTGAGACAGG	AACGTTGTAA	GACCTGTTGC	ACACAGTTGG	GCAGGTTGTG	9660
CCCTGTACAG	AGGGATGGGC	TGAGAGGGGC	AGTTGCCTGC	ATCACCCATT	GCAGCAGACT	9720
GGAGGGAGTC	TGCTTGTTTG	TAGTTCCTCA	GTCAGCAGGG	GCCTTTTGTC	TTTCCTTCCT	9780
TTCTTTTTTT	TTTTTTTTTG	AGACGGAGTC	TCACTCTGTT	GGCCAGGCTG	GAGTGTAGTG	9840
GCACAGTCTC	GGCTCACTGC	AATGTCCGCC	TCCTGGATTG	AAGCGATTTT	CCTGCCTCAG	9900
CCTCCTGAGT	AGCTGGGATT	ACAGGCGCGT	GTCACCATGC	CCAGCTAATT	TTTGTATTTT	9960
TAGTAGAGAT	GGGGGTTTCT	CCATGTTGAT	CAGGCTGGTC	TCGAACTCCT	GACCTCGTGA	10020
TCCGCCCACC	TCGGCCTCTC	AAAGTGCTGG	GATTACAGGC	GTGAGCCACC	ACGCCTGGCC	10080
AGCAGGGGCC	TTTTTTCTAA	TTTATATGAA	GACACCTAAT	TTATATGTGT	TAGCAAAGCC	10140
CTCCTGTTTA	TGCCTCACCT	CCTCCCCCGA	AGCTCATACG	GCAGGATGTT	CCTGAGAAAA	10200
TTGCCTCTTA	GAAGATAGAG	AGGAGATGCC	AAGCCTAAGT	TAGGCAGACT	CAGGAGGATA	10260
GGTCTGACCC	ACCCCCTGCC	ATTCCCCAGC	ACACTTGTGA	TTAATCTCCT	TGGCCAGAGC	10320
CAGGCAGAAC	ACCCTCGCGT	AAGAGATTTG	CCCCCCAGCC	CCGTCCCAGC	CCTCAGCTAG	10380
ACAGAAGATT	CCCTTTCCAG	AGAGGCTGCA	GAGCATGAGA	GCTCTTTCTG	TGTGCTTAAG	10440
GTCCCGTTCA	AAGGTGAGAA	AGTGAAGCTG	GTGCGGCTGC	GGAATCCGTG	GGGCCAGGTG	10500
GAGTGGAACG	GTTCTTGAG	TGATAGGTAG	GTGAGGGGAC	CCCACGGGAT	TGGCGGTGGC	10560
GGGGAACAGG	GTCCGGGACA	AGGCTGTGTT	GGGAACTGAG	CCATGAGAGT	ATTGAAGATG	10620
CTTGGTATAA	AATCACCCCTC	AAAACCAATG	ATCCGCAGAG	AAGAGGGGCA	CAGGTGTTGG	10680
CTCCAGGGAA	GGGCCAGGAG	TGGAAGCGGG	GTGCTGGGGA	CCCAGAGAGG	TTGCTGACAA	10740
CCATTGGCTG	GAAAGGAAGG	ATTCCAGAAA	GCGTGGGGAA	GGTCCAGGCA	GGAAAAGCGT	10800
ATGAATGCAG	GGTTCTGGGC	TAGAGAAAGT	ACTTCCCTTC	TTGGGGTCTT	GTGTTGCCTT	10860
TCCTGTGAAA	TGGGAACAGT	ATTATTAGCA	CTTACCTTGT	GGGCTGATAT	TGAGGAGTAA	10920

CTGGGACTTG TTTTGGGCA AGTGCTGAGC CATTGCTAAG ATTCCCCTTA CCCGTGCTTG	10980
TCCCTTGTAT TAAGGCACAA GGGCCCTTTG AAAAGAATTT TACCTGCTTT ATCAATTGAA	11040
AGGGATTAAG ACCTTGGGGG CCAACCCAAA ATAAACATGC GAACTTATTA TTTATAGGCT	11100
CCATGCACAC TTCGTAAAAC CTCCATGGTC CTACTGGTTC CTGATTACCT CCACTCAATG	11160
AGAGGCAATT CATTACTGAA TGAGCCATAA GCGCCTCTTA TTTCGAGAGG GGGATGGCAG	11220
GACTCAGTCG AGGAGAAGGA CCGCACCCAG GCAGCCTGGG CCCCTCGGCT CCTGTACTTA	11280
TTTACTGCTG GGTACTTCCT AGCCCAGCAT GTAATTACTG GTTCGTTTTCAG TCATTGTTTT	11340
AGTAAATGTT TCTTGGGCAC CTACTACATA GGAGGCACAG GTCAAGGCAC TGGGGATATT	11400
CTTTCTACCC ACCCCCTCCC TCCCTACACT GTGATTAGGG ACTGACCGAT C	11451

(2) INFORMATION POUR LA SEQ ID NO: 3:

(i) CARACTERISTIQUES DE LA SEQUENCE:

- (A) LONGUEUR: 1834 paires de bases
- (B) TYPE: acide nucléique
- (C) NOMBRE DE BRINS: double
- (D) CONFIGURATION: linéaire

(ii) TYPE DE MOLECULE: ADN (génomique)

(xi) DESCRIPTION DE LA SEQUENCE: SEQ ID NO: 3:

ATTTTTTTTT TTTTTTTTGA GACGGAGTCT CACTCTGCCA CCCAGGCTGG AGTGCAATGG	60
CGCGATCTTG GCTCACTGCA ACCTCCGCCT CCCGGGTTCA AGTGATTCTT CTGCCTTAGC	120
CTCCTGAGTA GCTGAGACTA TAGGTGCCCC CCACCACGCC CAGCTAATTT TTGTATTTTT	180
ATTAGGACGG GGTTCACCA TATTGGCCAG GCTGGTCTCG AAATCCTGAC CTTGTGATCC	240
CCCCACCTCG GCCTCCCAA GTGCTGGGAT TACAGGTGTG AGCCATTGCG AGCAGCCCAG	300
AACTCAATTC TTAACCTTTA AAGTATGATG AGAAGAAGGA TCAAGCCCTC ACCAGCCCAT	360
TTAAGGAGTT TAGGCTCACT CTTGAGGATG TGAGAAGTCA TTGCTATTGG GTTTCACACT	420
GAGGTTAACA GGTGAAGTCA GCATTTTGGT AGTTCACAGC AGCTGCAACT CTTTGTATTT	480
CTCTGATACC TCCTGTCCCA ACCTACATCA GGCCTTCCCT TCTTCCTGCT TCCTTAATTC	540
CTCCATTTTC CCACCAGATG GAAGGACTGG AGCTTTGTGG ACAAAGATGA GAAGGCCCGT	600
CTGCAGCACC AGGTCACTGA GGATGGAGAG TTCTGGTGAG TCCAGAACCC AGGAAGACCC	660
AGAAGGGTAA GGGTGGGGA GAGAGGGGA ATCTCAGACC TCAGTCCCCA GCTAAGGTTA	720
TCAGATTCCA GCCCTTGGGA GATCTTGGCT GTGTTCTCCT CCAGCCCAAG GCCCAGCAAG	780
GATGAGGTTT TGAGAGGAGC CTTCCAGGCC ACAGGGACAA TGAGCCCAGG ACCAGGCCAA	840
CATGACATGG CTCTTGCCCT CTGTGTGCCC CTCCGCCACA CACTCTATTC CAGCCACAGG	900
CACCCTGGCC TTAGCACAAT TCTTTTCTGA GCCTAGGAAG CTCCACTTAC CCTGATCTTC	960
CAACGTCAAC CTCACCCTCT CTCAGGTTGT TTCTATTCAG GCTTCAAGTC TCAGCTTAAG	1020
GAGAATTTTC AAGTCTCAGC TTAAGGAGAG CCCCCTAAGT TCCCCGAGGA CTGGGATTAA	1080
TTTATGATGC TCATCACCCT TAAAATTGTT TGCTTAAGCC GGGCGCGGTG GCTCACGCCT	1140
GTAATCCCAG CACTTTGGGA GGCCGAGGTG AACGGATCAC GAGGTCAGGA GATCGAGAAC	1200

ATCTTGGCTA ACACGGTGAA ACCCTGTCTG TACTAAAAAT ACACAAAAAA AGTAGCCGGG	1260
CGTGGCAGCG TCGGCCTGTA GTCCTAGCTG CTGGGGAGGC TGAGGCAGGA GAATCACTTG	1320
AACCTGGGAG GCAGAGGTTA CAGTGAGCCC AGATTGCGCC ACTGCACTCC AGCCTGGGCG	1380
ACAAGAGAGA CTCTGTCTTG GAAAAAAAAA AAAAAATGTG GTCTTAGTTT AATGTCAAGG	1440
GAAAGGTTTT GGGTGTTTTT ATTACTTTAT TTTTATTTA AAAACTATAA TAGAGACGGG	1500
CCTCGCTATA TTTCTCGGGC TGGTCTCAA CTCCTGGGCT CAAGCGGTCC TCCCACCTTG	1560
GCCTCCCAA ATGCTGGCAT GTGGGCCTGG TCAACATATG GGACCCCAAC TCTACAAAAA	1620
ATTTTAAAT TAGCCAGATG TGGTGGCGTG TGCCTGTAGT CCCAGCTACT TGGGAGGCTG	1680
AAGCAGGGGG TCACTTGAGC CCAGGAGGTT GAGGCTGCAG TGAAGTATGA TTGTCGTTCA	1740
CTTTTCTTCT GAACGTGAGA TTAAGTGTAG TCAGCAATTT GGCTTAGGAT TATTTATTCA	1800
GAATTTTAA CCGTCACGTT GCGGCAAACC AGGT	1834

(2) INFORMATION POUR LA SEQ ID NO: 4:

(i) CARACTERISTIQUES DE LA SEQUENCE:

- (A) LONGUEUR: 14664 paires de bases
- (B) TYPE: acide nucléique
- (C) NOMBRE DE BRINS: double
- (D) CONFIGURATION: linéaire

(ii) TYPE DE MOLECULE: ADN (génomique)

(xi) DESCRIPTION DE LA SEQUENCE: SEQ ID NO: 4:

AGGAGGTGGA GGTTCAGTG AGCCAAGATC ATGCCACTGC ACTCTAGCCT GGGCAACAGA	60
GCGAGACTCT GTCTCAAAAA ATACACACAC ACACACACAC ACACACACAC ACACACACAC	120
ACACACATAT ATATACACAC ATATATATAC ACACACATAT ACACACACAC ACGTCTGTAT	180
ATATATGTGT GTGTGTATAT ATACACACAC ACACTATTCT ATATATTCTT GTAGAGCTAT	240
GTGTGTCTCC TGTGCTATTG AGCATGAGCC CTTTTTTTTT TTTTTTTTTT TTGAGACAGA	300
GTCTCACTTT GTCGCCCAGG CTGGCATAGA ATGGCGCAAT ATCGGCTCAC TGCAACCTCC	360
GCCTCCTGGG TTCAAGTGAT TCTCCTGCCT CAGCCTCCCA AGTAACTAGG ATTACAAGTG	420
CCCGCCATAA TGCTCAGCTA ATTTTGTAT TTTCAGTAGA GATGGGGTTT CACCATGTTG	480
GCCAAGCTGG TCTCAAACCTC CTAGCCTCAG GTGATCCACC TGCTCAGCC TCCCAAAGTG	540
CTGGGATTAC AGGCATGAGC CACAGCACCC TGGTGAGCAC TAGAGCTTAT TTCTTCTATC	600
TAAGTGTATT TTTGTATCCA TTAGCCACCC TCTTTTCATC CTCCCTCTC CTTCCCTTCC	660
CAGCCTCTGG TAACCACTGT CTGCTCTCTA CTTCCATGAC ATATGCTTTG TTTAGCTCT	720
CACATATGAG TGAGAGCATG CGACATTTAT CTTTCTGGCC CTGGCACATT TTTGAATCAT	780
TGTTAGAAAA GATGATGGTT TGGAGTAGAT ACATCAGAAG TGACAGCGTT TGCCCTAAAA	840
AGGAAAGACA GGCTCCTCTG GGACCCTGAC CAAGTTCCTG TGAAGTATTT TATTATTGTG	900
CTGTGTTAGT CCTGGGGTCT TCCGTTCCCA GCCCTCCTCA CCTGCTCCCA TATGGCTCTC	960
TCTCTTCTTC CAACCTCTCA GGATGTCCTA TGAGGATTTT ATCTACCATT TCACAAAGTT	1020
GGAGATCTGC AACCTCACGG CCGATGCTCT GCAGTCTGAC AAGCTTCAGA CCTGGACAGT	1080
GTCTGTGAAC GAGGGCCGCT GGGTACGGGG TTGCTCTGCC GGAGGCTGCC GCAACTTCCC	1140
AGGTGGGAGA TGCTCTTGAT GGGGGGAGGG TCTAAGCCGA AAAAGTTCCA GGCAGAAGAA	1200

FIG. 8D/1

GCCTAACTAG TGCTTATTAA GTCTCTCTGT TCCAGACGTC CACTATCTTA TTAAACCTTC	1260
CCTGTTTTAC TGAGAAGGAA ACCACCATGC TGAGAAGTTT GCAATAGGGA GCTGGGTAGC	1320
AACTTTGGAA GCAGGAACCTT GTGGGAACAA TGCAGATGCT GCTTGGACTT ACGATGAGGT	1380
TATGTCCAGA TAAGCCCATC CATCTTTTGA AAATACCCTA AGTGAAAAGT GCATCCAATA	1440
TGCCTAACCC CCCAAACCTC ATAGCTTACC CTGGCCTACC CTCAAACATT GCTCGGAACC	1500
CTTGACCTTA AGCCTAAAGT TGGGCCAAAT CATCTAACTC CAAAGCCTAT TTTACAAAGA	1560
AAGTTGTTGT AATATCTCCA TGTAACCTAC TTAATACTTG TACCTAAAAA GTGAAAAACA	1620
AGAATGGTTG TACGGGTACT CGAAATCCAG TTTCTACTGA ATGTGCATCT CTTTCACATT	1680
GTAAAGTTAA AAAATTGTAG CCGAACCATC CTAAGTCAGG GACTGTGAGT ACTGTGTCAG	1740
TAACAGTAAG GGCCTATTG GAGAACCAAG TTAGCAGCTG CTGCAATAGT TCAAGTCAGA	1800
GATGATGAAA ACCTAGACCA AGTCAGTAGC AGCAGAGATG GAGGGGAGAC AGCAGATTTA	1860
GGGAGAGCAT ATTGGGTGAT GTAGGGAAGG AAGAAGAATG ATGTCAAGAT TCCCAGTTGG	1920
GGACCTGACA ACATTGCAAC ATAAGACACA CAAGAAGATC GGGTGGGTGG CTCATGCCTA	1980
TAATCCCAGC ACTTTGGGAG GCAGAGCCAG GAGGATCACT TGAGCCCAGG AGTTCAAGAC	2040
CAGCACAGGC AACATAGTGA CACCTCATCG TTACCCAAAA TAAAAAAAAA AATGAGGTGG	2100
GAGGATTGCT TGAGCTCGGG AGGTTGAGGC TACAATAAAC TGTGATCATG CCACTGCACT	2160
CCTGCCTGGG TGACAGAGTG AGACCCTGCC TCAAAAAAAAAA AAGACACACA AGAGAAAAAT	2220
ATCAGCGTGT TGTTTGTTTT TGGTGGAGTT AATTGTGGGG TTCTAGGGAA AGGAATTTAG	2280
CTTGGGACAT GGAAAGTTT AGGTTCTGT AGAGTGTCCT AGTGAAGATT TGTAATAGAG	2340
CATCGGATGC GCATATTAGA TGGCACTTGG TGATATGATA AGAACTCAAA AAATATTTGA	2400
GGAATAAAGG AAAGAAGAGG CCAGACGTGG TGGCTTATGC CTGTAATCCC AGCACTTTGG	2460
GAGGCTGAGG CAGGCGGATC ACTTGTGGTC AGGAGTTCTG GACCAGCTTG GCTAACATGG	2520
TGAAAACCCA TCTCTACTAA AGATACAAAA ATTAACCGGG GATGATGGTG GGTGCCTGTA	2580
ATCCCAGCTA CTTGGGAGGC TCAGTCAGAA GAATCGCTTG AACCCAGGAG GCGGAGGCTG	2640
CAGTGAGCCG AGATCGCGCC ACTGCACTCT AGCCTGGGCA ACAGAGCCAG ACTCCGTCTC	2700
AAAAAAAAAA AAGTGAGAGA GATTGAGGCT GGGATATATG GCTCAGGCAT CATGCGCGTG	2760
TAGGGGGCAG TTAAAAAGCA GAAGTAAGAA AGATTGCCTA GGGAGGCAGG AAGGGTGAGG	2820

TGAGAGGAGA	AGAGGCCAG	GACCAGATTC	TAGTCACCAA	CAGCGTTTAA	GGGGCAGGTA	2880
AGGAAAACAA	AACCATCAGC	AAAGACTGAG	AATGAAAGCC	CAGAGAGGAA	GGAAAAGCCA	2940
CACATACAAT	CAGTACAGCT	CCATCTGAAT	AAAGGTAGCG	CCCCCCCCC	CCCAAATCAT	3000
TAGAGAAATG	CCTGATTCCG	TTTTCTGTGG	ATTTTCTCTA	AGAACCTAGA	TGTGGGGAAT	3060
AGAAATAAAT	GGTTCCTCT	GTCTCATCCC	CTCCCTGCCC	TCTGAGAGGA	AGCTGTGATT	3120
GCGTGCTCCC	TTTCTGGGGG	TGCAGATACT	TTCTGGACCA	ACCCTCAGTA	CCGTCCGAAG	3180
CTCCTGGAGG	AGGACGATGA	CCCTGATGAC	TCGGAGGTGA	TTTGCAGCTT	CCTGGTGGCC	3240
CTGATGCAGA	AGAACCGGCG	GAAGGACCGG	AAGCTAGGGG	CCAGTCTCTT	CACCATTGCC	3300
TTCGCCATCT	ACGAGGTGTG	TAGTCCTGAT	TGGCTCCAGC	CCAGGAAACA	TACTTTCCCA	3360
GAGAGGACGC	TTCCAGGGGC	TTCTAGAGGG	GCCCTCTGCT	TCCTCAATAC	CAGTGACCCA	3420
CAGAGCTCCT	GGTATCAGGA	CCACTTGTGT	TTGTAACAAG	CAAAAATAC	CAGGGGGGGC	3480
ATTAGAGAGG	CAGTGGAGCG	GGCCTGGCAG	AACAGGTGCC	TGGGGGTCAG	GCTTCCGCAT	3540
GCGGGCTGCA	GTTGCTGGCA	TTGCCTTCCG	CAGGCTCCTC	ATCCTCATTC	ACATCTGAAG	3600
CATCTTCCTT	TCTGTTTCTT	CTCAAGGTTT	CCAAAGAGGT	ATAGCAGCAG	CAGCGGCCAG	3660
CAGTTGTGTG	CAGCACTACC	CAGGGGGGCC	CGAGTCTGTC	TGTGGCTCGT	CGAGAAGCTT	3720
CCTGGTGGGG	TTTGTGGGCA	GGACTTGTGA	TAGGAGAGGG	CCTTGCCTGT	TGTTATTTCC	3780
CACTTGCAGA	GCAGGTTGCC	TCAGGGCATT	GCATGACCCA	TGACTACCAC	CCCCAGGATG	3840
TGCACTTTCT	CCCTCGCACC	AGACACTGCA	CGTCACACAC	ATGCCTTTGC	ACACTCACCC	3900
TCCTCCACGC	TTACAGCCAC	ACACACAGTC	ACACAGACGC	GTTCTGAGGG	TGGCTGCCCC	3960
CTTGGGATGG	AGGAATCACT	TCCCTCAGAA	CCCAGCCAAG	TCCTCTAGGC	CTCCTTGGGG	4020
GTCCTTCCAG	CCTGAGGGGC	TTCGGAGCTG	AGGACAGCTG	TTCTGGTAAG	TGTCCCTGAG	4080
TGTGGGGATG	ACACATTTCC	ATTCACTCTG	AATCACAACA	GAAAAGGGAA	GAGGAATTGA	4140
GGTAGGGAGC	CTATTTAACC	CTTGGGAGTC	GGGAAGTAGG	GAGGTTGAAA	CTGTGACATG	4200
GGTGACCAGG	GAGTTGGGAA	GGGACCCTTG	GAGGTGGCTG	TGGCAGGACA	GGACGTTCTT	4260
CCCAGGGGGC	TCATGTGCCC	TGGGCTCTCC	CCATCTCTCA	GATGCACGGG	AACAAGCAGC	4320
ACCTGCAGAA	GGAATTCTTC	CTGTACAACG	CCTCCAAGGC	CAGGAGCAAA	ACCTACATCA	4380
ACATGCGGGA	GGTGTCCCAG	CGCTTCCGCC	TGCCTCCCAG	CGAGTACGTC	ATCGTGCCCT	4440

CCACCTACGA GCCCCACCAG GAGGGGGAAT TCATCCTCCG GGTCTTCTCT GAAAAGAGGA	4500
ACCTCTCTGA GTGAGTGCTG GCCCAGCTTT CCCACGTGTT TCTAAAAGCT CACATGGCCC	4560
ACTCCAGAGG TTGAAGGCAT GAGGCAGCTA GACACGTCTC CTCCAGGGTC CTTCTGCTGC	4620
TCCTGAGCCA CTGGCCACAT TACCCCCATT CATTGATTCA TCCATTCTGT GATATTTATT	4680
GAGCACCTAC TATGTTCCAG GCACTGTCCT AGGCACTAAG GATAGAGTAG TGAAGTAAAC	4740
AGAAAGAAAT CCCTGCCTTC ATGGAGCTTA ATATTCTAAC ATGAGACAAT AATGGATAGG	4800
AAAAACATAT GTAGCATGTT AGATTTGGAG AGGTGATATG GAGCAAAAAT AAAGTAGGGA	4860
AGAGGGATAG GAGGTGTTGG GGATGCTTGA AATTTTAGGT TAGCATGGCC AGGAAAGCCA	4920
CATCCTGTCC CTGGCCACCA CAGATGAGCT CATAGCCCCT GCCACTCTGA TCTCTGTCCT	4980
TGGAAGATGC ACCAGGTCCA TGGGTAGGTG GCTGGGTCAT GCCTTTGGGG GGCTCTGAGC	5040
AATACTAACA AGAACCTGCG TGCCTGGGCT TGGCTGTCCG GGATGGTGCT GACATGGGGC	5100
TGGTTCCTGG GGTGGGGTG TTCCAGGGGT TCTCTAGAGG CTGGTTCTGG CTTGGCTGCC	5160
AGGAAGCCGT GCACCAGAGC AAACCGTCCA CGGGCCTCCT GCTTGCTTCT GGTGACACTG	5220
AGACCCACA TGTCTGTATT CCTCACAGG AAGTTGAAAA TACCATCTCC GTGGATCGGC	5280
CAGTGGTGAG TGGTTTAGAT CTTCTGTGCG AAAAGTCCAG AGGGTCCCCT TCCCTGACCA	5340
TGCAGGGGAC AGATGGTGCA GGGGAGAATG GGCACCTGGCA GAGGGAATGG GAGTCTGGGC	5400
TGTGCTGAGC AGTCCCTCCT TGGCACTGCA AATCCTACTT TGGCATGGCC AGAAGTAATC	5460
GGCCTTAAGC ACCGGGGGCC ATTGAGGCAG TTCAGGGGCT GGGAAATATG GAAGAGGGTC	5520
CTGGAAGGA GAAGCAATTT GAACAATCGG AGGGAACAAG GCCACAGGAA GGGATGACAA	5580
GAGCCGCAGC GAACACTGGA TTCTGAGACT GGATAACATT GGATTTTACA CATAGAGAAA	5640
AGAAAGTAAG CTGGTGCCGG ACCTGGTGTT GACACTTGGA TCCTCCACTT ACCAGCGGGG	5700
TGACCTGGAC AATTTCTGTA ATCCCTCTCA CTCAGTTTCC TACTCAGTAA AACGGGGATG	5760
ATAATGTGCC TTGCAAGGCT TTTGTGAGGC TTCATCAATG AGGTGATGTA TGTGAAGTGT	5820
CTGGCACAGC ATGGGCACTC AAACAGAGGT GCTTTTTTAC ACTTTACACC TTACAAGGTA	5880
CTTTTCACAT GTGTCATCGC GATACTTGCA AGGTTGCTGA GAGGTAGATG GGGTTATAAT	5940
CCCTGGTGTT CAAGAAAGGA AGCAGAGGCT CAATGGGGTT GAATGACTTC TCTGAGTTCA	6000
CAGAGCTCAG TAAGTGGCAG GGTGGGAAC TCACATTCAG ACTCTCTGAC TCCAGACTTA	6060

GGTTTTTCCG	CACCTCCACG	CTGAGGCCAG	CCCCAGGCAG	TGAGAAGCCC	AAAGTCCGAA	6120
GCACAGAGTG	CTGTGTGTTG	GGCTCTGTGT	GTTGAGGAGT	CTTGTGACTG	CCTTGGGGCT	6180
TTGGGCTGTA	GTCAGCTGAC	AGTCCTTTGT	GCTCTGTGGG	GATGACGTAG	GCCAATGGGA	6240
GGACAAATGC	CCCTCTGAAC	TGTCTTCTGG	GCAGTGACAG	TCATGGTCAT	AATCCTGACC	6300
CTGAGCCAGT	GCCAGGTCTC	CAAGTGCCTT	CTGAATGACC	ACAGGCGATT	GGTTTTAGTG	6360
GTAGGTGCGT	GGGGATCTGT	TCTGGTCATC	TGGATGCTGG	TCATCGGGTG	CAGTATTGAT	6420
CAGGACCTGC	AAACCCAAAA	GCTTATGGGA	GCTGGCACGT	CACGTGAGTA	GAGCAGGCAG	6480
GTGCAGGGTT	TTTGATGTCC	CTGCACTGAC	ACAGTTGTCT	GCAGTTCTCC	AATTTGACAT	6540
TTGGGCTCCA	GTGTCGAGGG	TCAAACAAGG	AATTTTGGGG	CGTGGGCCAA	ATCTGGGAAG	6600
ACACAGGGAG	CAGGGCCCTT	TGGCTCAAGC	TGATAGTTGC	CGCAGGGATT	ACCAGGCCCA	6660
GGGCAGCCTG	CCACAAGCTG	GGGCTTTTAC	CAAAGAAAAT	CTCCCTATGT	TAAATGCTTG	6720
CTCAAAAATT	TTTAAAAAAT	ATTCTGTAAG	TCAAAATCCA	TTGTTAGGTC	AGTTTGAGAG	6780
AGCCATGTTT	TTGGTGTTTT	AGTAACCAAT	TTCATTTTTT	TATTATTTAT	TTATTTGTTT	6840
ATTTTTGAGA	CGGAGTTTCA	CTCTTGTCAC	CCAGGCTGGA	GTGCAATGGC	ATGATCTCAG	6900
CTCACTGCAA	CCTCCGCCTC	CCGGGTTCAA	GCAATTCTCC	TGCCTCAGCC	TCCTGAGTAG	6960
CTGAGATTAC	AGGTGCCCAC	CATCAGCCTT	GGATAATTTT	TGTATTTTTT	AGTCGAGATG	7020
GGGTTTCACC	ATGTTGGCCA	GGATAGTCCT	GAAGTACTGA	CCTCAGATAA	TCCGCCCACC	7080
TCAGCCTCCC	AAAGTGCTGG	GATTACAGGC	ATGAGCCAGC	ACGCCC GGCC	ACCAATTTC A	7140
TTTTTTAAAA	AAGGAAGAAA	GAAAACCTTA	GCCAGAAGAT	CTTTTTCCTT	GCCATATGCA	7200
GTAAGAGTAG	ATTATAAAAA	CAAAGTCAGA	GCAGTCACTG	GTGTCTGGGC	ATGGAGGAGA	7260
AAGAAGAATT	CTCTTCTCCC	TTCACCCTCC	ATGCCCCTTT	TTGGCTCCAT	GTGATTGAGA	7320
TTTCTGGACC	CTGGAGCCCC	ACCCCAAGCT	AAAGACCAGG	ATACAGGGAA	GCCACAACCA	7380
CTGGCGGTTT	TGAGAACTTA	CTTTTCACTT	ATTCTGCATT	TACTGTTTCC	TTTTCTTATG	7440
CAGAAAAAGA	AAAAAACCAA	GGTAGGTGTG	TGGGTAGAGA	GCATGAAGTG	TGTGTACTCA	7500
TGCATATGTA	TGTGCATGCA	TGTGAAGTGT	GCATGTGTGA	GCTCATATGC	ATCCATGCAC	7560
CAGACTTGCC	TCTTCCTCCC	CCTCCTTCCT	GAGCTTCTGC	TGGGGCCGAG	CGTGCCAGTAA	7620
TGACAACTAC	GATTTGCTGG	GGGAAGGCTA	CGTGCCAAGC	ACTCTTTTAG	GTGCTTTCCA	7680

TGATTAATTC CTTCTCACA ACAGCCCTAT GAGATTAGTA CTATAACTAT CCCCATTTTC	7740
AGAGGGAGAA AAGGTACAGA CTTGACTAAC TTGCCCAAGG CCACACAGCC AGAGAGGGGC	7800
AGAGCCAGTA CTTAGAGCCA GGCAGTCTGG GTCCAGAGTC CGTGTCTGA ACCACAAGAG	7860
GCCATCATAC GCCATCAGAT TTGGTGCTAG CATTCTGGT GGTGCCTGGT GGTGATGGAT	7920
CCATCACAGG GGTCTCCAG GTACTGGTGC TGGCCCAGAC CAGAGCTGAC ACTCCTCAGG	7980
CACTACCACA TTCCAGGCAC TGTGCTTGGG GTCAGTCCCT CTCTTTTTTT TCCCCCCAA	8040
TTATAACAGT ATCTACAAAG TAGGTGCTGT TATTTTTCCC CTTTCACAGG TGAGATAGAC	8100
TCAAAGAAGT GAACTTGCCC AAGGAACAGA ACTAATGAGT GGGGAAAATG GAACTGGAAG	8160
CCATGTCTGT TTAATCCAAA ACCTGTGTTT CTTGCCCTCT TTCTCTGATG CCAGCCCCCT	8220
ACACTTCAAG GCCTGTGTTG TCCAGACCCA CACTCGGGCC TGCCAGTGTG TGCCTGGCAG	8280
GGATGCTCCA TGGCCACACC ATATCCATCC TACACATCCC CCCTCAGACT GTGACCTCCA	8340
TTTGCTCTGG GATCCCCACA AGCTTCAGCT GCTTGAGCAA GACACTGCTT AGAAGGCAGA	8400
GCAAGCCAAG GCCTCTGGGG CCTGCTGGGA GCCAAAGCTG GGGAGCCGTT TCCAGGGGTC	8460
TATCTGCTTG AGCTGTCCTA GATGAGCAGC ATGGAAGGGC AGTGGTGCAT GAGTCCAGGC	8520
GGGCTGCTTT TCTGCTCCGA GAGGCTCTGC CTGCCAGTT GTTCTCTGCA TTGCAGCCTC	8580
AATCCCCACA GCCTTGCCCT CCCCCGGCTT TCCCTACAGG TGCACCGCAT CCACAGTGTT	8640
GGCACCATGC AGCAGCCGCT CTCGGTCCTT TTCATATCCT TGTCAGTTGC ACGAGCATGT	8700
CTTGAAAATA TCCCTTGTTT GTGTAGCATC TTAAATGTTT TTGCAGTATG ATTTTGCATT	8760
CAGTATCTCA TTTGATCCCC ACAAGAGCCC TATGAGGAGG GAAAGCAGAT TTTACCATTA	8820
AAGGATGAGT AACTGAGGC CAGAGAGGAT ATTTTGGTT TTTTGTGAGA CAGTCTCACT	8880
CTGTACCCA GCCTGGAGTG CAGTGGCTTG ATCTTGCTC ACTGCAAGCT CCACCTCCCA	8940
TGTTACACC ATTTTCTGCT CTCAGCCTCC CAAGTAGCTG GGACTACAGG CACCCACCAC	9000
CACACCCAGC TAATTTTTTT GTATCTTTAG TAGAGATGGG GTTTCACCCA GTTAGCCAGG	9060
ATGGTCTTGA TCTCCTGACC TTGTGATCTG CCTGCTTCGG CCTCCTAAAG TGCTGGGATT	9120
ACAGGCGTGA ACCCCCCTGC CCGGCCAGAG AGGATATTTT TTAATGAGGG GCAGGGCTGG	9180
GATTCCAGCC CAGTGTCTG ATGGCTCACC CACTGACCAT TCCACTAATC CGTGTCTTTT	9240
TTCAATCTAA ACTTTCAGGG TTGTAGAGGT TCCTTTGAGG TGCCTCAGTA CTTCATGGT	9300

GATGTGGGGT	CTGAGGGCCA	AGAGCTCTGT	TCTCATTAAT	CAGAGAAGCT	TGTGTTTTTA	9360
AAAACACCAT	GTTTACTGCA	GGAAATTTAA	TTGGACAGTG	TTTCCATCTG	GAAAAAAAAA	9420
AGTCTACAAA	ATACTTGACA	ATCACTGCAC	TAGATCATGC	TGCTTTTAGC	ATTCTTAGCA	9480
TTTCACGTGC	TGAGCTCTCA	ATACTCTACC	ATGAGGAGGG	ATGGAGTGGG	TATGAAAAGA	9540
TAAAGAACTG	AAGTCACACG	GCTTGTCACT	GGCAGAGATA	GAGCTTGAAC	CGAGGTTGAA	9600
GAGCTCCCGC	CTATTCTTTT	CCTCTTCTCA	CTGGATAAAG	CTGCTCCAAG	AGAGGTGCTG	9660
CCTCAGTGTC	CCTGTTCAGA	CTGTAATCCT	CCCTTCCTTC	CTGCCTCCTC	CCTCCTCTCT	9720
CCAGCCCATC	ATCTTCGTTT	CGGACAGAGC	AAACAGCAAC	AAGGAGCTGG	GTGTGGACCA	9780
GGACTCAGAG	GAGGGCAAAG	GCAAAACAAG	CCCTGATAAG	CAAAGCAGT	CCCCACAGGT	9840
GTCTGGGCAT	GTGGCATGGG	TGGGGTGGCC	AGCACGCTAC	AGGGGCTTCC	TATGCGCTTG	9900
GGATACACAG	GGGCTGGAGG	CTTCCCAGGA	GTTTGTCTTG	AACATCTGGA	GGTTTGAATT	9960
TGTCCCACTG	ACCTTTTCTT	TCAGCAAGTT	CCCCTGAAAT	TTGGGCTGCT	GCTTGGGTGA	10020
ATATCCCAGG	ATGGGGGTTC	CATTCTAGGA	GTGGACTGGC	AGGCTGAGCC	TCCCATGGAG	10080
CTGATCCAGC	CAGGATACAG	AGAAGGGGAG	GCAAAGGCTG	AGACAGAACC	AGCTTGAGAG	10140
CGGAGGCGCA	ACTCTTGTCT	CCTGGTGGCC	TTGAGCATTT	CACAATAGGG	GGATAAAGGA	10200
TAGGAGCAGA	AAAGTGGGGC	TGACTTCAGA	AATGGGGTCC	TCTAGAGCTC	ACGGGAGGGT	10260
GTTAGATTGG	AGTGGGAGCT	TAGTGGAGGT	GAGCCTTAGA	GGCAAAAGTC	TCCAGACCAA	10320
TCCAGGCCCC	CTCTTCTATC	CGGGGGCCCC	TCTTCTATCC	AGGGCCCCCTC	TTCTGTCTGG	10380
GAGCCCCTCT	TCTATCTGGG	GCCTCATGCA	GTGGGGCCTA	GGGGAGGTTT	TCTGAGGACT	10440
TGGCCTTGAT	GACAGGGTGG	CTGGAGGAAT	CAGAACGGTC	AGACCTTCTT	TGACCTGCGG	10500
GCACCTTTAG	TTGGAATGCT	CAGGCCTGGG	ATGGTGGAGG	GGGCTCTTGC	AGGTGGGGAC	10560
TGGGGTGGCG	GGGAGGAGGC	TGTATGGCCG	CCATATCTCC	TTTGGCTGGG	GGCGTCAGGG	10620
CTGGAGAGGT	GTGAAGAGTC	CCTGAGGCCT	CGATGCATCT	CACTCCAGCT	CACCAGGTCT	10680
GCATTTGCCC	GTCCCCAGCT	CCTGCTGCCA	CCCCCGGCCG	TTTTAGGCAC	TTGGCTCCCT	10740
TGGCCCAGAG	GAGCTTGCCT	CACAGGCCTG	TGCACCTCTG	ACCCCTGTGA	ACCAGTTTTT	10800
CTTTGTGCCT	CCACAGCCAC	AGCCTGGCAA	CTCTGATCAG	GAAAGTGAGG	AACAGCAACA	10860
ATTCCGGAAC	ATTTTCAAGC	AGATAGCAGG	AGATGTGAGT	ACCTCCAAGC	CCAGGACGCC	10920

CACAGGTGCT TCCTTCTCTC CTGGATTAACT TGCTCAGATT ACCAATTATT TCATTATTGT	10980
TTGGTAGAGG TCACTTTGGA CTTCCGGTGA GCCAGGGGAT GTGTGCGTAG CACACAAATC	11040
CACAAGCCCT TGAGTTTTGG ACTGCCACGT CTGCTGGGGG GCTCAGAGGC CTTTTTGCTC	11100
TGAGCTGCCC ACGGTGGTCC TGATAGCTGA GGTGCAGTAT CTGGCCCCCT GTCTTCCTCA	11160
GAAAAGCCCC AGCTTCCCAT GACATAATAG CACCGACAGG GATTTTACAA ACACAGCCAG	11220
GTGGAATTTG TTTTGCAAAG TGTCCGCGCC AGGAGCTGCT GTACTCCTGA ACCATGACCC	11280
TCCTCTCCCT TCCTCCTCAG GACATGGAGA TCTGTGCAGA TGAGCTCAAG AAGGTCCTTA	11340
ACACAGTCGT GAACAAACGT GAGTTGCTCA AACCAAATGG GGGTGGGGTG GGTGGGGAGT	11400
CCCGTTGTCT CAAAGCAGCT CCTCACTCTT CTCCATCCCC CCAGACAAGG ACCTGAAGAC	11460
ACACGGGTTC ACACTGGAGT CCTGCCGTAG CATGATTGCG CTCATGGATG TATCCTTCCT	11520
GCCGCCCCCT CCCGACCCTC TGTATCAGC CCACGGGGGC CAAGGCAACA TACAGGGTGC	11580
CCAGTCAGGC AAAGGGCCCT AATTGTGCC CAGGGAAACT TAAGGAGACC CTGATTCAGA	11640
ACATCTTGA TACTCGTCTG AAAGGGGTTG TTAGAGGCGG AAGGGGAGGA TGTGGGTTG	11700
TAAGTGCCT AACCCCTGTG CTTCTCTCAG GCCTGGGATC CTGCCCAAGC AAAAGTGCTC	11760
CTTAGGAGAG CGGCTCCTGG GTTACAGAGT AGGCGCAATC TCTGACTGGT GGTGGAGTGG	11820
AGGGGAGGGT TAAATAGTAC AACAGGGCAG TGGGTAGGAC AGCCCGGAGT CTCCTAGACC	11880
CTCCCTCAA ATCCAGGGG ATTTTGCTGT GTGCTGTGTA GCCCTGACCT CCTCCTCCA	11940
GACAGATGGC TCTGGAAAGC TCAACCTGCA GGAGTTCCAC CACCTCTGGA ACAAGATTAA	12000
GGCCTGGCAG GTGGGAAGAG AAAATGAAGC GTGGGAGTCA AGAATGGGGT TGATTTGGAG	12060
ATTCAGTGTG TGACCTCCAT CCTCAAATTT TCTATTGCCA GAAAATTTTC AAACACTATG	12120
ACACAGACCA GTCCGGCACC ATCAACAGCT ACGAGATGCG AAATGCAGTC AACGACGCAG	12180
GTGCTGAGAA GGAAGGGGTG TCAGGGATGT GGACCCGAGA CGGTGGGAGC AGGAATGGGA	12240
GGGACTAGC TACTAGGGCC CCACTAGAGA AGGAGAGGGA AAGGGCTTCT CACTTTCCT	12300
TCCCAGGTCA CAGAGTGTCC GAGAGGCAGG GAAAATAGAA GACAGGCCCA AGGCCTCCAG	12360
CTCCACGTCC ACCTCTAACA TGGTCCCCTC CACAGGATTC CACCTCAACA ACCAGCTCTA	12420
TGACATCATT ACCATGCGGT ACGCAGACAA ACACATGAAC ATCGACTTTG ACAGTTTCAT	12480
CTGCTGCTTC GTTAGGCTGG AGGGCATGTT CAGTAAGTGG GAGAGGGGGG CTGCCCTCTG	12540

CTCTCTTGCA	GGGGCAGTTG	TGGCAACAGG	CATCTCACCT	GATAATCTCC	AGTCTGCTCC	12600
ATCCAGGCTG	AACAAGGGCC	AATGACCTCT	TTAGGCCCAG	AATGGGATGG	CAAAGGGAGG	12660
GTTACTGGTG	ATTCTCTGCC	TGCACATCTT	TGTGCTGATG	AGGGACAGCA	CTGGGCACAC	12720
GGTCCTCTGA	GGGGAAGTTA	CAGTAGTAGA	GGCGGAGTGC	GCCTGTAACT	GGCCTCTGGC	12780
CTGTGCATTG	TTTCACAGGA	GCTTCTCATG	CATTTGACAA	GGATGGAGAT	GGTATCATCA	12840
AGCTCAACGT	TCTGGAGGTA	AAGCATAGGC	ACAGCACATT	CCCCCTACAC	ATTAAAACTC	12900
AAGGTGGAGG	GGTCAACGGG	GCGGACTGGA	CCCAGGGTGT	GCTCCTCATT	TCCACACAGT	12960
GGTGGAGGGA	AGGGATAGGA	ACAGAACATG	GAGGGAGGCT	CAGCAGGCTC	CCAGGACACA	13020
TGCACTTGAG	GCCCCAAAAGG	ACCTCTGCTC	CCCCAGTCAC	TTGATGCGGG	AAAACATGCA	13080
CCTTCTTAGG	GAAGATCTAG	GAGAAAGGAA	ACAGTAAGCC	ACTGCTTCTT	GGAAAACTTT	13140
CTGGGGGTCT	GACCTGCTGG	GACTGTTCCC	TTTCTCTTG	CCCCGTAAGA	TTCTAGGGC	13200
GGGGGGGGGG	GGGGGTCACT	CTTTTCTGAT	CTACATTCTG	ATCTTGGGAC	TTCTTTTCACT	13260
GGCTGCAGCT	CACCATGTAT	GCCTGAACCA	GGCTGGCCTC	ATCCAAAGCC	ATGCAGGATC	13320
ACTCAGGATT	TCAGTTTCAC	CCTCTATTTT	CAAAGCCATT	TACCTCAAAG	GACCCAGCAG	13380
CTACACCCCT	ACAGGCTTCC	AGGCACCTCA	TCAGTCATGT	TCCTCCTCCA	TTTTACCCCC	13440
TACCCATCCT	TGATCGGTCA	TGCCTAGCCT	GACCCTTTAG	TAAAGCAATG	AGGTAGGAAG	13500
AACAAACCCT	TGTCCCTTTG	CCATGTGGAG	GAAAGTGCCT	GCCTCTGGTC	CGAGCCGCCT	13560
CGGTTCTGAA	GCGAGTGCTC	CTGCTTACCT	TGCTCTAGGC	TGTCTGCAGA	AGCACCTGCC	13620
GGTGGCACTC	AGCACCTCCT	TGTGCTAGAG	CCCTCCATCA	CCTTCACGCT	GTCCCACCAT	13680
GGGCCAGGAA	CCAAACCAGC	ACTGGGTTCT	ACTGCTGTGG	GGTAAACTAA	CTCAGTGGAA	13740
TAGGGCTGGT	TACTTTGGGC	TGTCCAACCT	ATAAGTTTGG	CTGCATTTTG	AAAAAAGCTG	13800
ATCTAAATAA	AGGCATGTGT	ATGGCTGGTC	CCCTTGTTGT	TTGTTGTCTC	ACATTTAGAT	13860
ATCAGCCATG	CATGACTGAA	TGGCTTCCAA	TCATATACTC	ACCTATCACC	TACAAGAGAA	13920
CAATGAAAAA	CACACACAAA	AACAAAATCT	TGAATTTTGT	AATCATGCCT	ATTGCTATTT	13980
CTTGAGCATA	AGAATGGCTC	AGATACTTTC	CAAGACATAA	AAGGAAGGCA	GAGGAATAGT	14040
TGTTGCTGTA	AAAGACATCA	AGAATAAATG	GGGTCATGTA	CAACGGGAGG	GGCCGGTTAC	14100
CTGAATAATG	GAGTGGAGAT	TGAGCTATCC	TAGCTCCTCT	GCTCACTAAC	TGACCTGTGG	14160

CATGACCGTG GACAAAACCC TGAACGCAGC TGTTTGTTTG CTAAACTTCT CTGGACCATG	14220
GCCTGCGGCA TATCTATAGG CATCCTGTGT TTTCCACCCA GTTTCCTTCT TCCTCGCTAA	14280
GCCAACGTGG AAAGGGCTGG CCGTGAATAT GCAGACAAGG TAACGAAAGT AAACCGTCAA	14340
TTAGTAAAAG TACTTCATTT TCCTCTTGTA TTTGCTTCAT TCTTGCTTCA CAAAGTTACG	14400
AAGTCCACAG CTTTATACCA AAATGTAAGA AGGCTATTTG CTTATAAACA TTTTGAGTCA	14460
GGTGTCACTT GATTTCAATC TTCTAATCCA TATTCAATAT TAAAAAATCA GAAACCAAGG	14520
GTGCTGGAGC AGCTCTAGGG CATATATTTT TCTTAAATAG GAGAAAGATT TTCAACAGCT	14580
TTTCCTCCTT GACCCCCTCC TTTCCCAATT TATTGGGTC ACTACCTTGA ATTTAGAGTG	14640
AATCTGGGAA ATGTAGTCAC CAGG	14664

